Section 1.1

What is Statistics?

Section 1.1

What is Statistics?

Learning Objectives

At the end of this lecture, the student should be able to:

- State at least one definition of statistics.
- Give one example of a population parameter, and one example of a sample statistic.
- Be able to classify a variable into quantitative or qualitative, and as nominal, ordinal, interval or ratio.

Outline

- Definition of Statistics
- Population Parameter & Sample Statistic
- Classifying Levels of Measurement

Definition of Statistics

Statistics Helps Decision-Making

Concepts in Statistics

- What is statistics?
- Individuals and variables
- Examples of statistics, individuals and variables in healthcare

What is Statistics?

- Statistics is the study of how to collect, organize, analyze, and interpret numerical information and data.
- Statistics is both the science of uncertainty and the technology of extracting information from data.
- Statistics is used to help us make decisions. This is especially important in health care and public health.

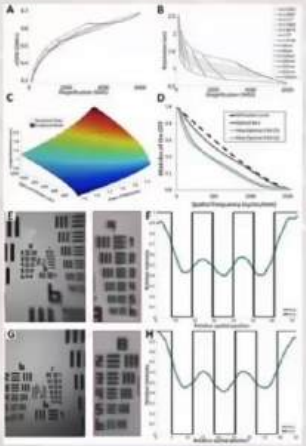


Image from Cybulski J, Clements J, Prakash M (2014). "Foldscope: Origami-Based Paper Microscope". PLOS ONE. DOI: 10.1371/journal.pone.0098781

Example: CDC & the Flu Vaccine

- During the year, the United States Centers for Disease Control and Prevention (CDC) *collects, organizes, analyzes, and interprets numerical information and data*
- They *extract information from the data* and *make decisions* about what to include in next year's flu vaccine.

Photo by Marcela

Individuals & Variables

Meaning Outside Statistics

- Individuals are people.
  - We expect 50 individuals at the graduation.
- A variable is a factor that can vary, possibly causing a problem.
  - The time the shop takes with my car is an unknown variable, and I can't predict it.

Meaning in Statistics

- Individuals are people or objects included in a study.
  - 5 individuals could be 5 people, 5 records, or 5 reports.
- A variable is a characteristic of the individual to be measured or observed.
  - The age of an individual person
  - The time an individual record was entered
  - The diagnosis listed on an individual report

A Few Examples

Individual	Variable
Kidney dialysis patient	Number of blood transfusions
Baby born to a mother who smokes cigarettes	Birthweight
Post-menopausal woman	Compliance with health screening recommendations
State with casinos	Rate of pathological gambling
County with hospitals	Medication error rate
Urban city	Rate of overdose deaths

Concepts in Statistics

- Statistics is used in healthcare and other disciplines to help aid in decision-making.
- Understanding statistics is necessary to understand certain processes in healthcare.

Photo by Jorge Franganillo

Population Parameter & Sample Statistic

Important Differences

Parameters vs. Statistics

- What is a population and what is a sample?
- Difference between population and sample data
- Population parameters and sample statistics
- Examples of parameters and statistics

What is a Population?

Definition

- A population is a group of people or objects with a common theme.
- When every member of that group is considered, it is a population.

Example

- *Theme:* Nurses who work at Massachusetts General Hospital (MGH)
- *Population:* List from Human Resources of every currently employed nurse at MGH

What is a Sample?

Definition

- A sample is a *small portion of the population.*
- It can be a representative sample.
- But it can also be a biased sample.

Example

- Only survey ICU nurses at MGH
  - Not a representative sample
- At least one nurse from each department
  - More representative sample

Population vs. Sample Data

Population Data

- In population data, data from every individual in the population is available.
- Entire population = census

Sample Data

- In sample data, data is only available from some of the individuals in the population
- Very commonly used in research studies of patients

Examples of Population Data

- Medicare Claims Dataset – has all the insurance claims filed by Medicare population
- United States Census (conducted every 10 years)

Examples of Sample Data

Photo courtesy of KayEss

- The Medicare Beneficiary Survey (MBS) is a survey of a sample of individuals on Medicare
- American Community Survey (ACS) conducted yearly by the United States Census Bureau

Statistical Notation

Total Population	Sample of Population
N	n

Parameter vs. Statistic

A parameter is a measure that describes the entire population

A statistic is a measure that describes only a sample of a population

Photo courtesy of Che/Wikimedia Commons

Photo courtesy of Sandstein/Wikimedia Commons

Examples of Parameters & Statistics

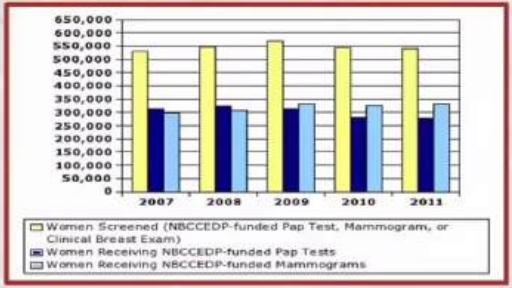
Parameter	Statistics
Mean age of every American on Medicare	Mean age of Americans on Medicare estimated using the MBS
The proportion of Americans addicted to cigarettes	The proportion of Americans in the Behavioral Risk Factor Surveillance Survey (BRFSS) who admit they are addicted to cigarettes
Actual voter turnout	The proportion of people in opinion polls who say they plan to vote

Don't Get Confused!

- When you hear a number on the television or radio – do they mention if it is a population parameter or a sample statistic?
- Clues that the number is a population parameter:
  - A dataset was used that encompasses the entire population (like Medicare)
  - Analyses were done by the government or on behalf of the government
- Clues that number is about a sample statistic:
  - It was from a study recruiting volunteers
  - The report mentions only surveying or measuring a sample of individuals

Describing vs. Inferring

- Descriptive statistics involve methods of organizing , picturing, and summarizing information *from samples and populations.*



Images courtesy of the Center for Disease Control and Prevention and the American Journal of Preventive Medicine.

Describing vs. Inferring

- Descriptive statistics involve methods of organizing , picturing, and summarizing information *from samples and populations.*
- Inferential statistics involves methods of using information *from a sample to draw conclusions regarding the population.*

Images courtesy of the Center for Disease Control and Prevention and the American Journal of Preventive Medicine.

Parameters vs. Statistics

- In statistics, it is important to properly identify measures as population parameters or sample statistics.
- Different types of data are used for parameters and statistics.

Map by Metric Maps.

Classifying Levels of Measurement

Four-level system

Classifying Variables

- Quantitative vs. qualitative data
  - Interval vs. ratio data
  - Nominal vs. ordinal data
- Examples of how to classify healthcare data

Photo courtesy of Toho-u

Quantitative vs. Qualitative

Quantitative is a numerical measurement of something

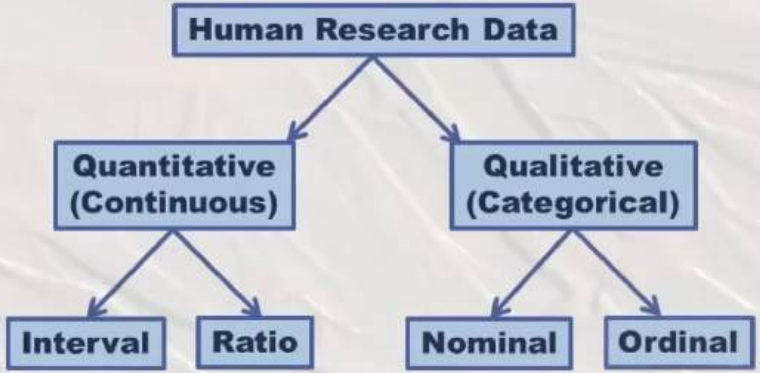
Photo courtesy of Xell/Wikimedia Commons

Qualitative refers to a "quality" or categorical characteristic of something

Photo courtesy of Ceedz/Wikimedia Commons

SUBSCRIBE

Four-level Data Classification



Examples of Quantitative Variables in Healthcare

- Time of admit
- Year of diagnosis
- Systolic blood pressure (SBP)
- Platelet count

Photo courtesy of US Navy

Interval vs. Ratio

Interval

- Quantitative (continuous) data
  - Differences between data values are meaningful
- There *is no* true zero

Ratio

- Quantitative (continuous) data
- Differences between data values are meaningful
- There *is a* true zero

Examples of Quantitative Variables in Healthcare

- Time of admit
- Year of diagnosis
- Systolic blood pressure (SBP)
- Platelet count

- Ratio = there is a true zero
  - If you are dead, you have a SBP = 0
  - Same is true with platelet count

SUBSCRIBE

Examples of Quantitative Variables in Healthcare

- Time of admit
- Year of diagnosis
- Systolic blood pressure (SBP)
- Platelet count

- Interval = no true zero
- "Time" cannot have a true zero
  - Time of admit = 8:09 am
  - Year of diagnosis = 1999

SUBSCRIBE

Examples of Qualitative Variables in Healthcare

Photo courtesy of US Navy

- Type of health insurance
- Country of origin
- Stage of cancer
- Trauma center level

Nominal vs. Ordinal

Nominal

- Nominal applies to categories, labels or names, and *cannot* be ordered from smallest to largest

Ordinal

- Ordinal applies to data that can be arranged in order in categories,
  - but the difference between data values cannot be determined, or is meaningless.

SUBSCRIBE

Examples of Qualitative Variables in Healthcare

- Nominal = Cannot be ordered
- Ordinal = natural order
  - Differences between levels is meaningless

- Type of health insurance
- Country of origin
- Stage of cancer
- Trauma center level

Conclusion

- Definition of Statistics
- Population Parameter & Sample Statistic
- Classifying Levels of Measurement

SUBSCRIBE

Learning Objectives

At the end of this lecture, the student should be able to:

- Define “sampling frame” and “sampling error”.
- Give one example of how to do simple random sampling, and one example of how to do systematic sampling.
- Explain one reason to choose stratified sampling over other approaches.
- State two differences between cluster sampling and convenience sampling.
- Give an example of a national survey that uses multi-stage sampling

SUBSCRIBE

Outline

- Sampling Definitions
- Simple Random Sampling
- Stratified Sampling
- Systematic Sampling
- Convenience & Multi-stage Sampling

Sampling Definitions

Terms you Need to Know

SUBSCRIBE

Concepts in Sampling

- What is a “sample”?
- Sampling frames, and errors in representing sampling frames
- Summary of definitions presented

Photo by Intel Free Pres

Sampling and Samples

Photo by Jayal Aheram

- We take a sample of the population because we want to do "inferential statistics"
  - We want to infer from the sample to the population
- Reasons not to measure the whole population
  - Impractical
  - Unnecessary

SUBSCRIBE

Sampling Frame

- List of individuals from which a sample is actually selected.
- "List" may be a physical, concrete list
  - List of students enrolled at a nursing college
- May be a theoretical list not made up yet
  - List of patients who will present to the Emergency Department today

Photo by Schizmo

SUBSCRIBE

Sampling Frame

Sampling frame is the part of the population from which you want to draw a sample.

Therefore, you want everyone from your sampling frame to have a chance of being selected for your sample.

SUBSCRIBE

Undercoverage

What is it?

- Omitting population members from the sampling frame

How can it happen?

- List of nursing students may not include everyone for administrative reasons
- People who present to the Emergency Department at night might be different than those in the day

SUBSCRIBE

Errors in Statistics

Fact-of-Life Error

- Sampling error
  - The population mean will probably be different from your sample mean
  - The population percentage will probably be different from your sample percentage

Error you want to Avoid

- Non-sampling error
  - Using a bad list.
  - Make sure that you pay careful attention that everyone in the population who is supposed to be represented in your sampling frame is in there!

SUBSCRIBE

Causes of Error

Fact-of-Life Error

- Sampling error – caused by the fact that, regardless of what you do, your sample will not perfectly represent the population.

Error you want to Avoid

- Non-sampling error – caused by poor sample design, sloppy data collection, inaccurate measurement instruments, bias in data collection, other problems introduced by the researcher.

SUBSCRIBE

Simulation

- A simulation is defined as a “numerical facsimile or representation of a real-world phenomenon.”
- It is a essentially working through a pretend situation to see how it would come out in the case it was real.
- That is why this course includes many simulations, or real-life examples.

SUBSCRIBE

Concepts in Sampling

- It is important to do your best to avoid non-sampling error
- This is achieved by making sure you do not have undercoverage when sampling from your sampling frame.

Simple Random Sampling

What it means

Simple Random Sampling

- What is simple random sampling?
- Two methods of randomly sampling from a list
- Limits of simple random sampling

Photo courtesy of Felix O

Definition & Example

Definition

- "A simple random sample of *n* measurements from a population is a subset of the population selected in such a manner that every sample of size *n* from the population has an equal chance of being selected."

Example

- You have a list of the *population* of students in a class.
- You want to take a *sample* of 5 (n=5).
- If you take a simple random sample (SRS) from the class list, it means all the different possible groups of 5 students you could pick from the list has an equal chance of being the sample (group) you actually pick.

SUBSCRIBE

One Method of SRS

- Number all of the individuals in the population with a unique number.
  - Like student ID number
- Put all the student ID numbers in a place from which you can draw randomly without looking (like a hat)
- Draw 5 ID's and use those students as your sample.

Photograph by US Army

SUBSCRIBE

Another Method of SRS

Photograph by US Navy

- Generate a list of random numbers as long as the list of the population.
- Randomly assign these numbers to the population in the list.
- Take the first 5 numbers (whoever gets assigned 1 through 5).

SUBSCRIBE

SRS Means Equal Chance of Being Selected

Photograph by Joe Mabel

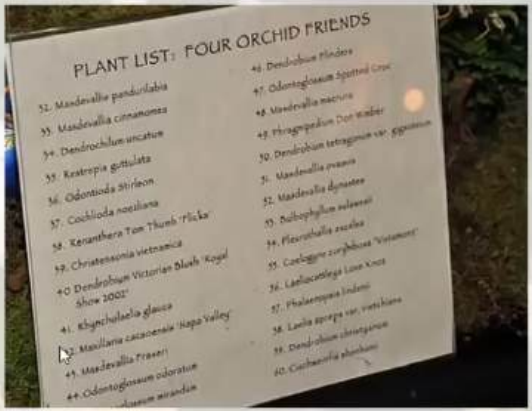
- *First method:* old-fashioned "hat"
- *Second method:* Electronic "hat"
- *In both methods*, all members of the population had an equal probability of being selected into the sample

Limits of Simple Random Sampling

- You need a list
  - You don't know who will present at the Emergency Department that day, how do you sample?
  - Okay when a list is available.
- You need a *good* list
  - Otherwise, you risk undercoverage
  - What if part-time students were not on the list?
  - Non-sampling error

Simple Random Sampling

- Characteristics of SRS
- Two methods of randomly sampling from a list
- Limits of SRS



Photograph by ShakataGaNai

SUBSCRIBE

Stratified Sampling

What it means

Stratified Sampling

- What is stratified sampling?
- Steps in stratified sampling
- Examples in stratified sampling
- Limitations of stratified sampling

Photo courtesy of Angie of Sawara, Chia-ken, Japan.

SUBSCRIBE

What is Stratified Sampling?

- First, the list is divided into groups, or strata.
- This is a way to make it so that there are certain proportions of groups in the final sample.
- Next, simple random sampling (SRS) takes place for each of the strata

Photograph by Wonderlane

SUBSCRIBE

Steps in Stratified Sampling

1. Divide entire population into distinct subgroups called strata.
2. The strata are based on a specific characteristic, such as age, income, education level, and so on.
3. All members of a stratum share this specific characteristic.
4. Draw an SRS from each stratum.

Examples of Stratified Sampling

- In a high school, sampling so many students from each of the grades (freshman, sophomore, junior, senior)
- In hospitals, sampling so many patients or providers from departments (different intensive care units)

Photo courtesy of Undergrounded199.

SUBSCRIBE

Limitations of Stratified Sampling

- Oversampling one group means your summary statistic is unbalanced
- It is not possible to do without a list beforehand (like with SRS)
- It also is hard because you have to split the list into groups ("strata") then SRS from the strata

Stratified Sampling

Photograph by Sciondriver

- Stratified = taken from groups.
- Several steps are involved.
- Useful if necessary to make all strata equal, or to sample from groups that are small in the large population

Systematic Sampling

What it means

Systematic Sampling

- Systematic sampling can be done with or without a list!
- Systematic sampling is best described through the steps one takes to do it

Photo courtesy of Immanuel Giel.

Steps in Systematic Sampling

1. Arrange all individuals of the population in a particular order.
2. Pick a random individual as a start.
3. Then take every *kth* member of the population in the sample.
   - "*kth*" means "every so many".

Examples of Systematic Sampling from a List

- Take out a list of classes available next semester.
- Pick a random number that is small – like 3. Go to the third class.
- Pick another random number – like 5. Pick every 5th class after that.

Photo courtesy of Hobvias Sudoneighm

Characteristics of Systematic Sampling

- You *cannot* do this when there is a pattern to the data (boy/girl/boy/girl)
- You *can* do systematic sampling in a clinical setting, where you do not know who is going to come in that day

Photos courtesy of Nils Simon and US Army.

Systematic Sampling

Photograph by US Navy

- Systematic sampling is easy to do with or without a list.
- Just pick a random starting point, then pick every *k*th individual.

Cluster Sampling

What it means

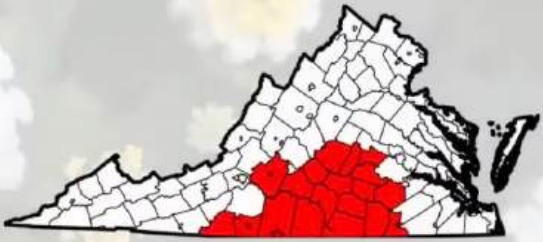
Cluster Sampling

- Why use cluster sampling when you could use stratified, systematic or simple random sampling?
- Because the problem is in a particular geographic location

Photo by Simisa.

Why use Cluster Sampling?

- The problem is localized to a particular location
- In cluster sampling, we begin by dividing the map in geographic areas.
- Then we randomly pick clusters, or areas, from the map. We take all the people in the cluster.



Map by Dbenbenn and Makeemlighter.

Problems with Cluster Sampling

- Sometimes, the people located in a cluster are all similar in a way that makes the problem hard to study.
- If cancer rates are high all over the clusters, it's hard to see if a geographic location is causing higher rates.

Photo courtesy of Hansueli Krapf

Cluster Sampling

Photograph by Born Londoner

- Cluster sampling is used when geography is important in sampling.
- The map is divided into areas, and all the people in a particular area are sampled.
- Biased toward type of people living in the area.

Convenience & Multi-Stage Sampling

What it means

Convenience Sampling

- Convenience sampling can be used under low risk circumstances
- What ice cream is the best from the restaurant next to the hospital?
- However, often results are not reliable

Photo by Managementboy.

What is Convenience Sampling?

- Using results or data that are conveniently or readily obtained.
- Can be useful if not a lot of resources allocated to the study.
- Use an already-assembled group for surveys.
- Ask patients in the waiting room to fill out a survey, or students in a class.

What are the Problems with Convenience Sampling?

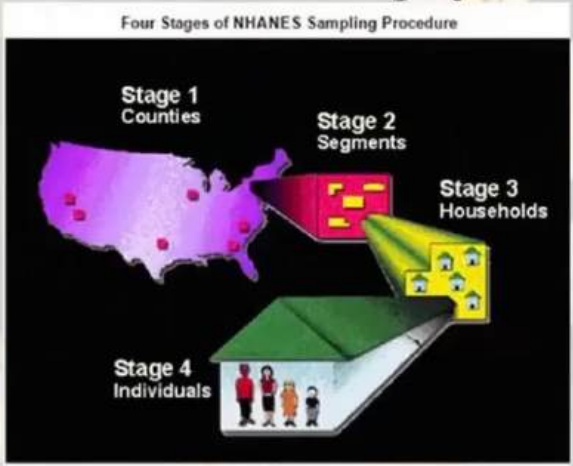
- There is a bias in every group.
- Often miss important subpopulations (what stratified sampling addresses).
- Results can be severely biased

Photo by Fanch Gaume.

Multi-stage Sampling

- Combination of sampling strategies layered in stages.
- Example:
  - Stage 1: Cluster sample of states (two census regions)
  - Stage 2: Simple random sample of counties (from each state)
  - Stage 3: Stratified sample of schools (urban/rural)
  - Stage 4: Stratified sample of classrooms

National Health and Nutrition Examination Survey (NHANES)



Courtesy of the Centers for Disease Control and Prevention.

Convenience & Multi-Stage Sampling

- Avoid using convenience sampling unless the question is low risk
- Use if the only type of sampling possible under the circumstances
- Also used when resources are low
- Multi-stage sampling usually used in large, governmental studies.

Photograph by FBI Buffalo Field Office

Conclusion

- Sampling Definitions
- Simple Random Sampling
- Stratified Sampling
- Systematic Sampling
- Cluster Sampling
- Convenience & Multi-stage Sampling

Section 1.3

Introduction to Experimental Design

Learning Objectives

At the end of this lecture, the student should be able to:

- State the steps of conducting a statistical study.
- Select one step of developing a statistical study, and state the reason for this step.
- Name one common mistake that can introduce bias into a survey, and give an example.
- Explain what a lurking variable is, and give an example.
- Define what a completely randomized experiment is.

Introduction

- Steps to Conducting Statistical Study
- Basic Terms & Definitions
- Avoiding Bias in Survey Design
- Topics in Randomization

This Section

- Review the steps to conducting a statistical study
- Define vocabulary terms
- Examples provided from healthcare

Basic Guidelines for Planning a Statistical Study

1. State a hypothesis.
2. Identify the individuals of interest.
3. Specify the variables to measure.
4. Determine if you will use the entire population or a sample.
    - If you choose a sample, choose a sampling method
5. Address ethical concerns before data collection.
6. Collect the data.
7. Use descriptive or inferential statistics to answer your hypothesis.
8. Note any concerns about your data collection or analysis
    - Make recommendations for future studies

Hypothesis & Variables

- *Hypothesis:* Air pollution causes asthma in children who live in urban settings
- *Individuals:* Children in urban settings
- *Variables:* Air pollution and asthma

A person whom goes by the name Imagere

Sampling, Ethics & Data Collection

Photo courtesy of US Army Africa.

- Either collect data or use existing dataset
  - Can use a government dataset for population measures
- Can collect data from a sample for estimates
  - Need to choose sampling approach
  - Will need consent if legally found to be "human research"
  - May need consent from parents to collect data about children

Sampling, Ethics & Data Collection

IRB

Photo courtesy of US Army Africa.

- lect data or use existing

a government dataset
tion measures
ta from a sample for

oose sampling
oach

- Will need consent if legally found to be "human research"
- May need consent from parents to collect data about children

Census vs. Sample

In a census, measurements or observations from the entire population are used.

Photo courtesy of Che/Wikimedia Commons

In a sample, measurements or observations from *part* of the population are used.

Photo courtesy of Sandstein/Wikimedia Commons

Experiment vs. Observational Study

Experiment

- A *treatment or intervention* is deliberately assigned to the individuals
- The *purpose is to study the possible effect of the treatment or intervention* on the variables measured

Photo courtesy of US Marines.

Experiment vs. Observational Study

Experiment

- A *treatment or intervention* is deliberately assigned on the individuals
- The *purpose is to study the possible effect of the treatment or intervention* on the variables measured

Observational Study

- Observations and measurements of individuals are taken
- However, *no treatment or intervention is assigned* by the researcher

Examples of Experiment vs. Observational Study

Experiment (with OS)

Women's Health Initiative

Observational Study (OS)

Nurses Health Study

Replication

- Studies must be done rigorously enough to be replicated.
- Replicating the results of observational studies and experiments is necessary for science to progress.

Photo courtesy of Neils B.

Review

- Several steps need to be followed in order when conducting a statistical study.
- It is necessary to determine the type of study, and make other study-related decisions.

Avoiding Bias in Survey Design

Important Concepts

Bias

- Surveys can provide a lot of useful information
- However, it is important that all aspects of survey design and administration minimize “bias”
- Several considerations should be made

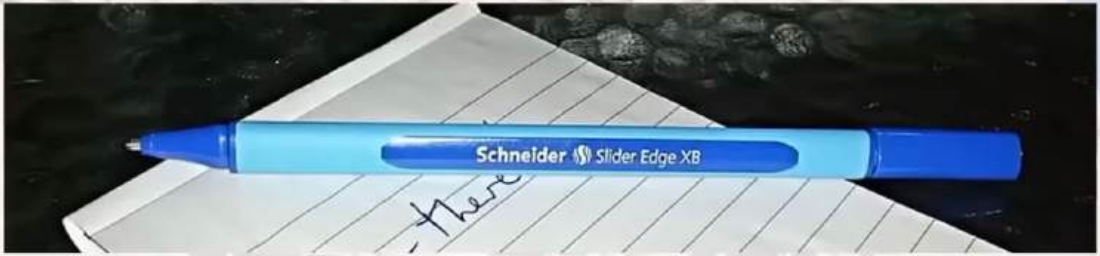


Photo by Revital9

Non-response & Voluntary Response

- If many people refuse your survey, the people who do complete it are likely to have a biased opinion.
- There may be a reason they do not complete your survey that has to do with how they feel about your survey topic.

Photo by Shay of Belfast, Northern Ireland.

Truthfulness of Response

- Respondents may lie on purpose
  - If asked a question that is too personal
  - If asked a question too hard to think about
- Respondents may lie inadvertently
  - May not remember if asking about something that happened a long time ago
  - May have “recall bias” influenced by events that have happened since original event

Hidden Bias

- Question wording may induce a certain response.
  - How long have you been using Software A?
- Order of questions and other wording may induce a certain response.
  - Do you agree with Obamacare?
  - *More people have health insurance than ever before.* Do you agree with Obamacare?
- Scales of questions may not accurately measure responses
  - Do your feelings always fit on a scale of 1 to 5?

Interviewer Influence

Photo by UK Department for International Development.

- This is important with in-person and phone surveys
- Best to have interviewer from same population as research participant
- All verbal and non-verbal influences matter

Vague Wording

- Avoid vague terms used in a survey.
  - Instead of asking if a person waited “a long time” in the waiting room, ask the number of minutes.
- If you must use vague terms, include grounding language.
  - Where 10 is extremely important, and 1 is not at all important, how important is having a controllable lifestyle to you in your future career? A controllable lifestyle is defined as one that allows the physician to control the number of hours devoted to practicing his/her specialty.

Lurking Variable

- "Lurk" means to sneak around behind the scenes
- A "lurking variable" is a variable that is associated with a condition, but may not cause that condition.
- For example, we know that having more education increases income. However, people of the same education level do not all make the same income.
- Lurking variables are sex and race. In the US, according to the Department of Labor, these variables can decrease a person's income at the same level of education.

Lurking Variables Cause Confounding

- Current studies show that why women and African Americans make less money on the whole is not explained by *fewer of them working, or fewer of them getting degrees.*
- Early studies were *confounded by these lurking variables.*

Photo of Jackie Lacey, Los Angeles County District Attorney, by Neon Tommy.

Final Note on Bias

- Survey results are important for improving healthcare and the progression of science.
- It is important to pay special attention to avoiding bias in the design and conduct of surveys.

Topics in Randomization

Definitions & Terms

Randomization

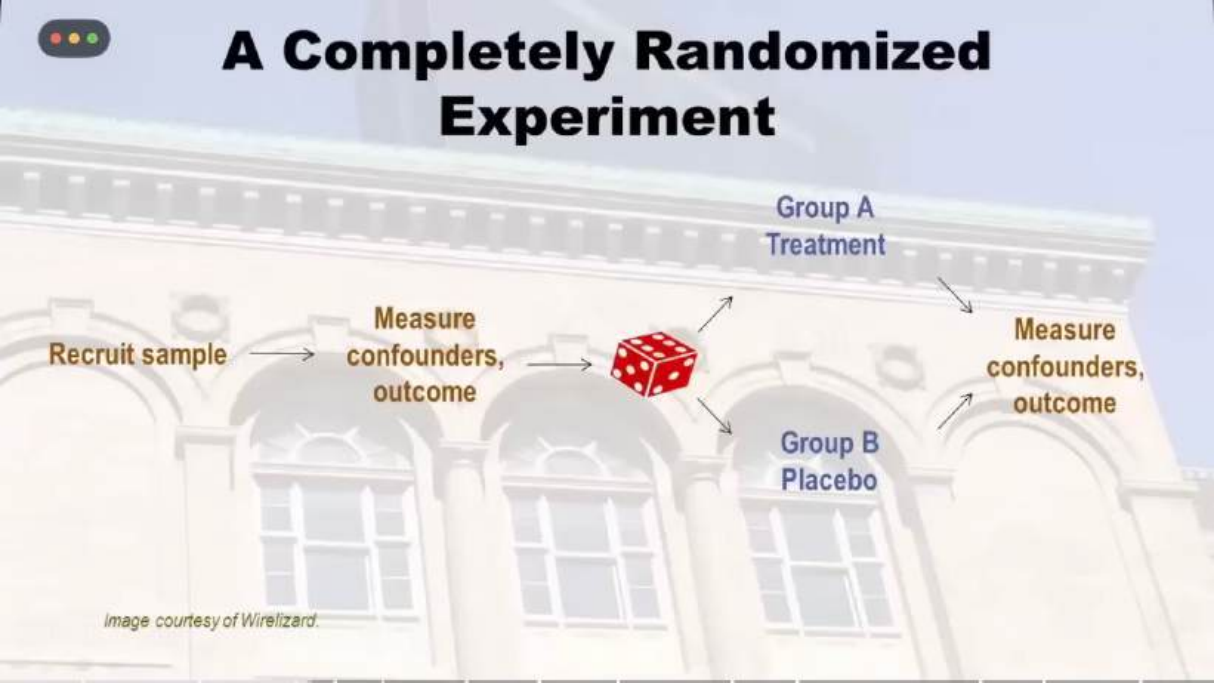
- Steps in a completely randomized experiment
- Placebo and placebo effect
- Blocked randomization
- Blinding

Why Randomize?

- Randomization is used to assign individuals to treatment groups.
- This helps prevent bias in selecting members for each group.
  - It distributes "lurking variables" evenly

Image courtesy of Svjo.

A Completely Randomized Experiment



Placebo & Placebo Effect

- Placebo effect occurs when there is no treatment, but participant assumes s/he is receiving treatment and responds favorably.
- The placebo is given to a control group, which receives the placebo (or attention control if treatment is not a drug).
- Used as a control or comparison group.

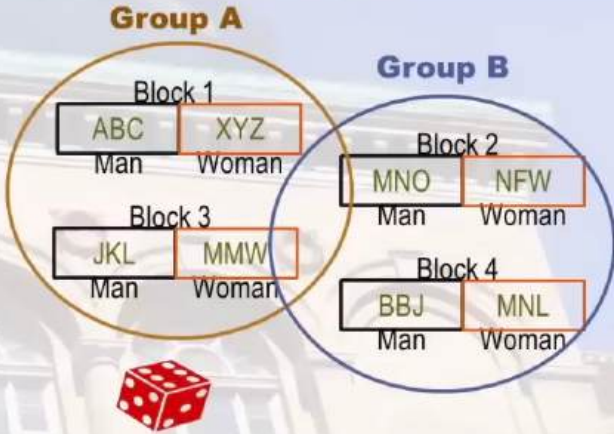
Image courtesy of the National Institutes of Health

Blocked Randomization

- If you want men and women equal in two randomized groups, create "blocks" with two slots – one for a man, and one for a woman.

Blocked Randomization

- If you want men and women equal in two randomized groups, create "blocks" with two slots – one for a man, and one for a woman.
- As people come in and enroll in the study and you measure them, assign them to blocks.
- Then, randomize the blocks. You will get a paired couples in each group!



Blinding

- Blinding is where a person (participant, research staff) is deliberately not told of a treatment assignment in a study so s/he is not biased in reporting study information.
  - *Example:* A participant is blinded to treatment or placebo.
  - *Example:* A study radiologist may be blinded to treatment group when looking at images during study procedures.
- Double-blind means study staff and participant do not know treatment assignment.

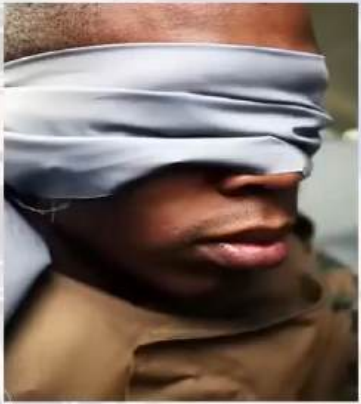


Image courtesy of U.S. Marines

Randomization & Bias

Photograph by US Navy

- Randomization is used to reduce bias in an experiment.
- Blocked randomization can even out groups.
- Blinding further prevents bias
- The placebo effect is necessary to take into account.

Conclusion

- Steps to Conducting Statistical Study
- Basic Terms & Definitions
- Avoiding Bias in Survey Design
- Topics in Randomization

Photo by photosteve101

Chapter 2.1

Frequency Histograms

&

Distributions

Learning Objectives

At the end of this lecture, the student should be able to:

- State the steps for drawing a frequency histogram.
- Name two types of distributions and explain how they look.
- Define what an outlier is.
- Say one reason why you would make a frequency histogram.
- Define relative frequency and cumulative frequency.

Introduction

Photo by BlairSmith66

- Review of frequency histograms and relative frequency histograms
- Description of five common distributions in statistics
- Explanation of outliers

What is a Frequency Histogram?

Charting the Frequency Table

Frequency Histogram

- Explain what a frequency histogram is
- Describe the steps to drawing a frequency histogram
- Explain relative frequency histogram

Photo by Rego Korosi

What is a Frequency Histogram?

- It's a specific type of bar chart made from data in a frequency table.
- Frequency histograms and relative frequency histograms.
- The purpose of the chart is to identify the "distribution" of the data.

Photo by Loqueveo

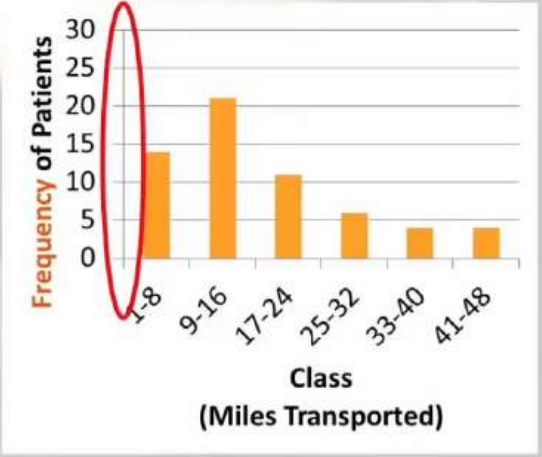
Steps to Follow to Draw a Frequency Histogram

1. Make a frequency table.

Class Limits	Freq-uency	Relative Freq-uency
1-8 miles	14	0.23
9-16 miles	21	0.35
17-24 miles	11	0.18
25-32 miles	6	0.10
33-40 miles	4	0.07
41-48 miles	4	0.07
Total	60	1.00

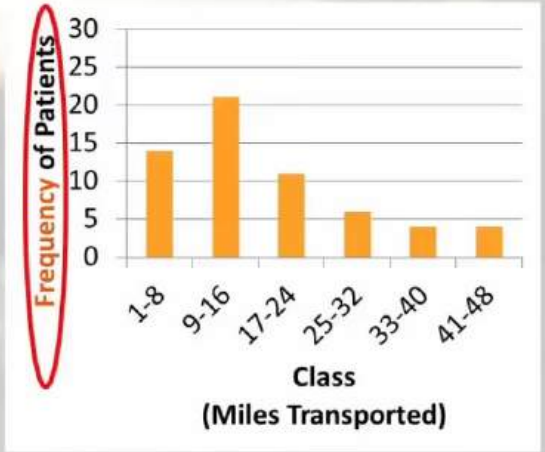
Steps to Follow to Draw a Frequency Histogram

1. Make a frequency table.
2. Draw a vertical line for the y-axis.



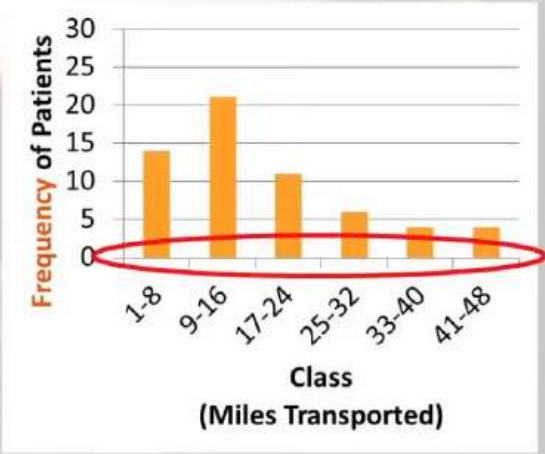
Steps to Follow to Draw a Frequency Histogram

1. Make a frequency table.
2. Draw a vertical line for the y-axis.
3. Write "Frequency of _____" along the y-axis.



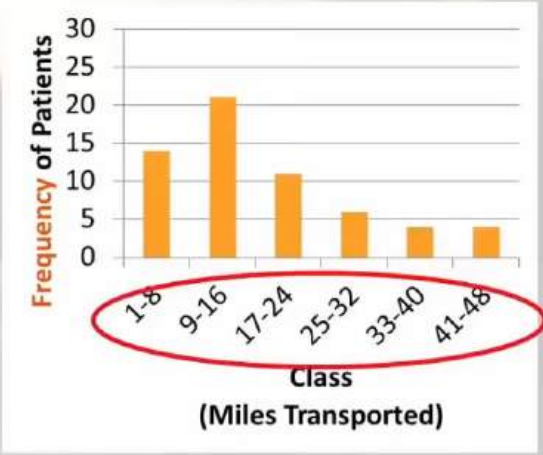
Steps to Follow to Draw a Frequency Histogram

1. Make a frequency table.
2. Draw a vertical line for the y-axis.
3. Write "Frequency of _____" along the y-axis.
4. Draw a horizontal line for the x-axis.



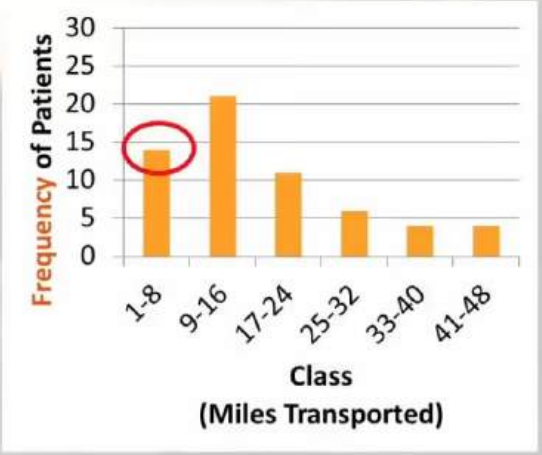
Steps to Follow to Draw a Frequency Histogram

1. Make a frequency table.
2. Draw a vertical line for the y-axis.
3. Write "Frequency of _____" along the y-axis.
4. Draw a horizontal line for the x-axis.
5. Write the classes below the x-axis and label them.



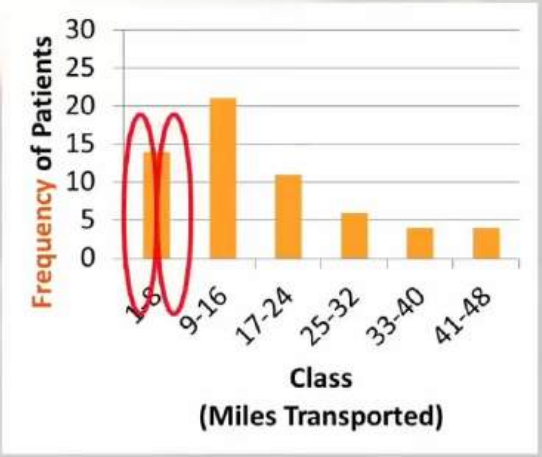
Steps to Follow to Draw a Frequency Histogram

6. For the first class, find the frequency in the table. Look for it on the y-axis and draw a horizontal line.



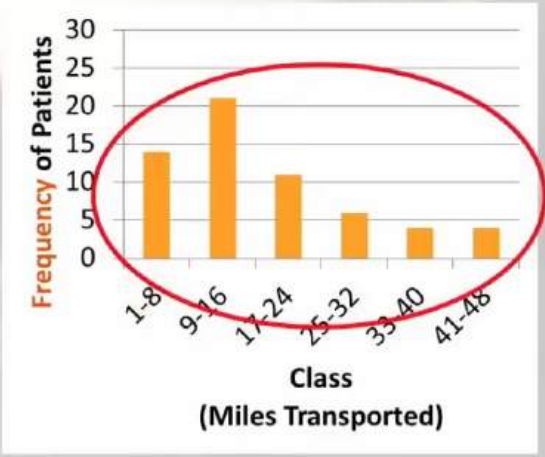
Steps to Follow to Draw a Frequency Histogram

6. For the first class, find the frequency in the table. Look for it on the y-axis and draw a horizontal line.
7. Draw two vertical lines down to make a bar.



Steps to Follow to Draw a Frequency Histogram

6. For the first class, find the frequency in the table. Look for it on the y-axis and draw a horizontal line.
7. Draw two vertical lines down to make a bar.
8. Repeat for all the other classes.
9. Color in the bars



Relative Frequency Histogram

- In the relative frequency histogram, the relative frequency goes on the y-axis.
- The chart looks takes on a similar pattern.
- Relative frequency better for comparing two populations or two samples.

Frequency & Relative Frequency Histograms

Photo by Gabriel Miguel Gutierrez Valenzuela

- After making a frequency table, it is important to also make a frequency histogram and/or a relative frequency histogram.
- These are used to reveal the “distribution” in the data

Understanding Distributions

Frequency Histograms Reveal Distributions

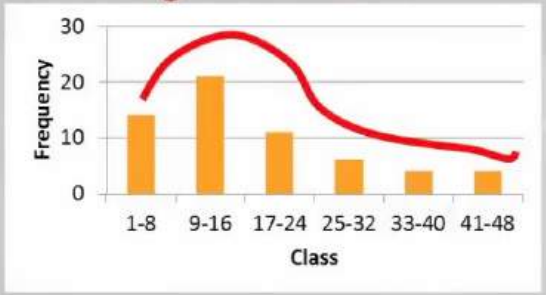
Distributions

- Define distribution and why it is important to know the distribution
- Describe outliers and how they can be found using histograms
- Example cumulative frequency and ogives

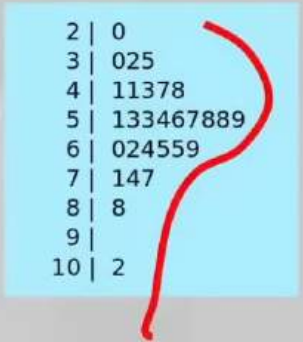
Photo by Keith Hall from UK

What is a Distribution?

- It is the shape that is made if you draw a line along the edges of a histogram's bars.

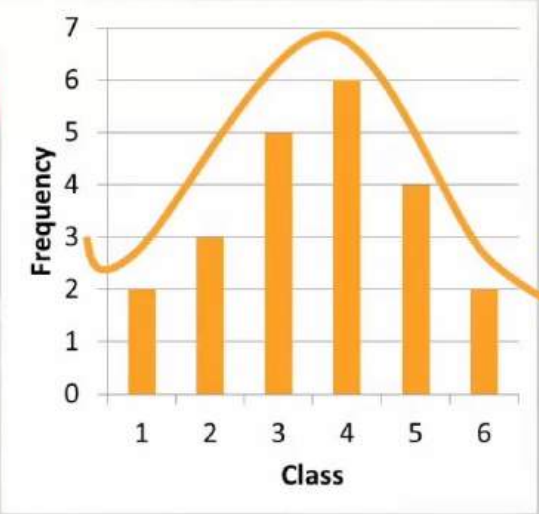


- A stem-and-leaf of the same data will make the same shape on its side.



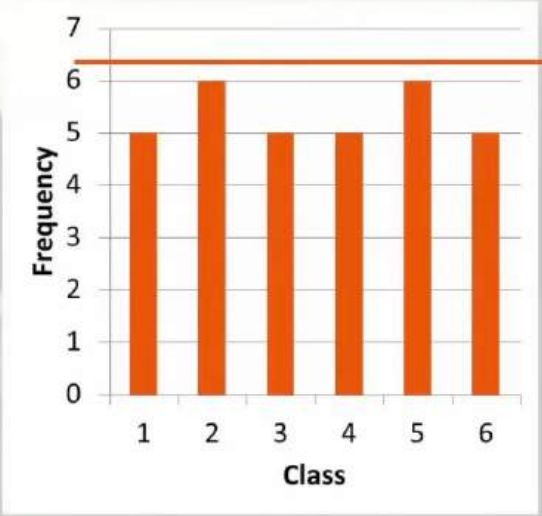
5 Main Types of Distributions

1. Normal distribution (also called mound-shaped symmetrical)
2. Uniform distribution
3. Skewed left distribution
4. Skewed right distribution
5. Bimodal distribution



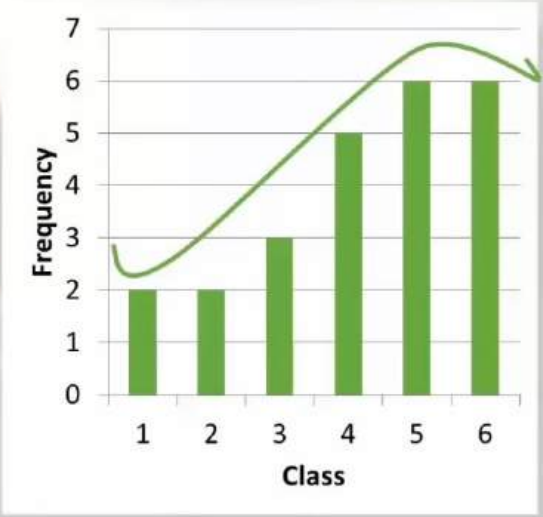
5 Main Types of Distributions

1. Normal distribution (also called mound-shaped symmetrical)
2. Uniform distribution
3. Skewed left distribution
4. Skewed right distribution
5. Bimodal distribution



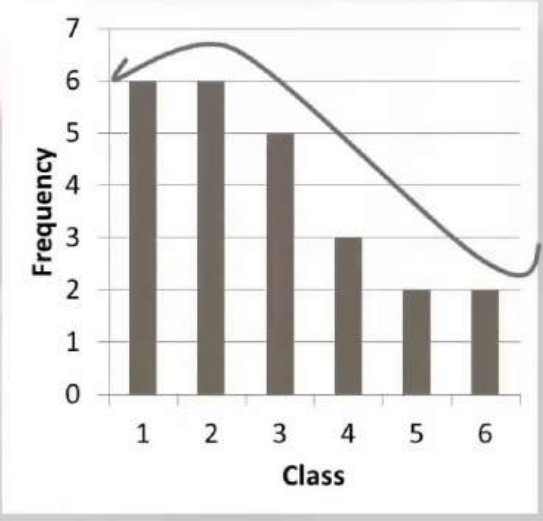
5 Main Types of Distributions

1. Normal distribution (also called mound-shaped symmetrical)
2. Uniform distribution
3. Skewed left distribution
4. Skewed right distribution
5. Bimodal distribution



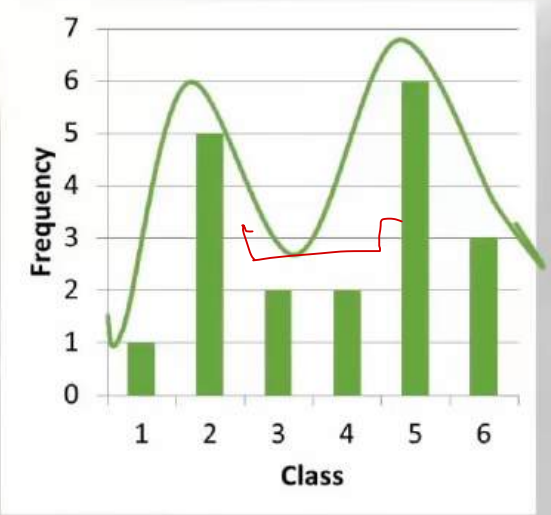
5 Main Types of Distributions

1. Normal distribution (also called mound-shaped symmetrical)
2. Uniform distribution
3. Skewed left distribution
4. Skewed right distribution
5. Bimodal distribution



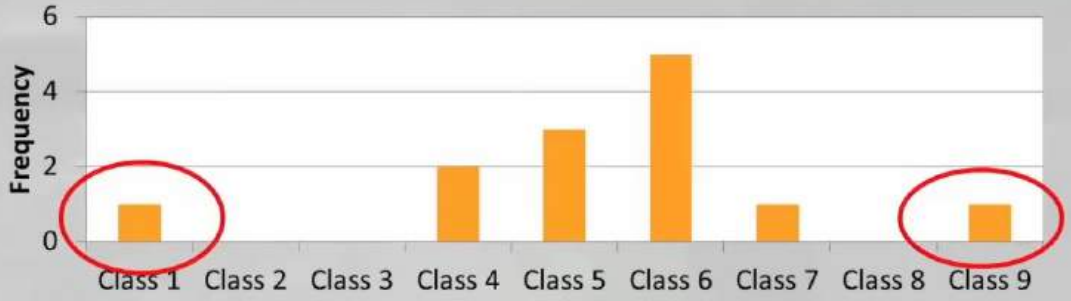
5 Main Types of Distributions

1. Normal distribution (also called mound-shaped symmetrical)
2. Uniform distribution
3. Skewed left distribution
4. Skewed right distribution
5. Bimodal distribution



Outliers

Outliers are data values that are “very different” from other measurements in the dataset.



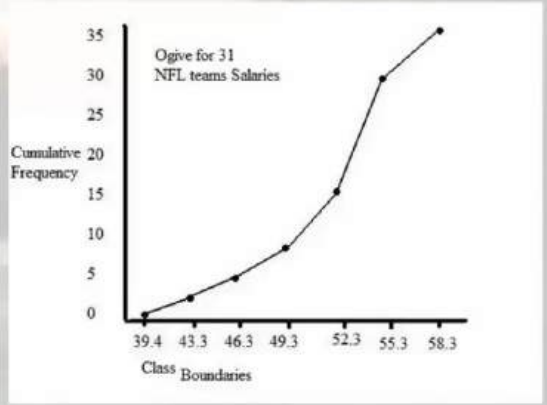
Cumulative Frequency

- In "cumulative frequency", you add up all the classes before the class you are on.
- The first class is always the same as the frequency.
- Each cumulative frequency is equal to or higher than the last one.

Class Limits	Freq-uency	Cumulative Frequency
1-8 miles	14	14
9-16 miles	21	14+21=35
17-24 miles	11	35+11=46
25-32 miles	6	46+6=52
33-40 miles	4	52+4=56
41-48 miles	4	56+4=60
Total	60	60

Chart of Cumulative Frequency: Ogive

- Classes along the x axis, and cumulative frequency along the y-axis
- Because cumulative frequency goes up from class to class, the ogive line always goes up to the top frequency.



From JLW87/Wikimedia Commons

Distributions

Photo by Lariob

- There are 5 main types of distributions used in statistics.
- Histograms and stem-and-leaf displays are used to look for outliers.
- An ogive is a chart of cumulative frequency.

Conclusion

- The purpose of the histogram is to reveal the distribution
- Stem-and-leaf displays also reveal the distribution
- Knowing the distribution is important in statistics

Photo by Mark Dixon

Chapter 2.2

Other Graphs

Learning Objectives

At the end of this lecture, the student should be able to:

- Describe a case in which a time-series graph would be appropriate.
- Explain the difference between what would be graphed on a bar graph vs. a time-series graph.
- Describe the type of data graphed in a pie chart.
- List two considerations to make when choosing what type of chart to develop

Introduction

- Explanation of time series graph
- Explanation of bar graph
- Explanation of pie chart
- Review of graphs – when to use what type of graph

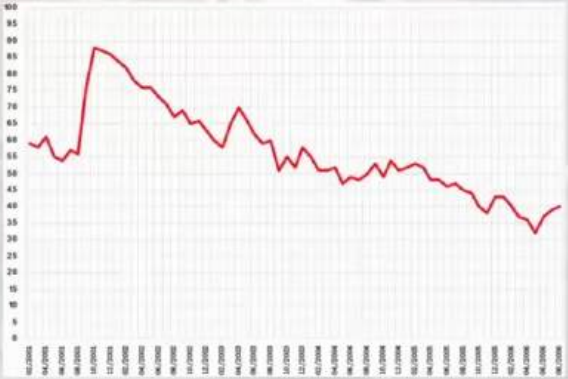
Image courtesy of US Air Force

Time Series Graph

Graphing Data Over Time

Time Series Graphs

- Define time series graph and time series data
- How to plot a time series graph
- Purpose and interpretation of time series graphs

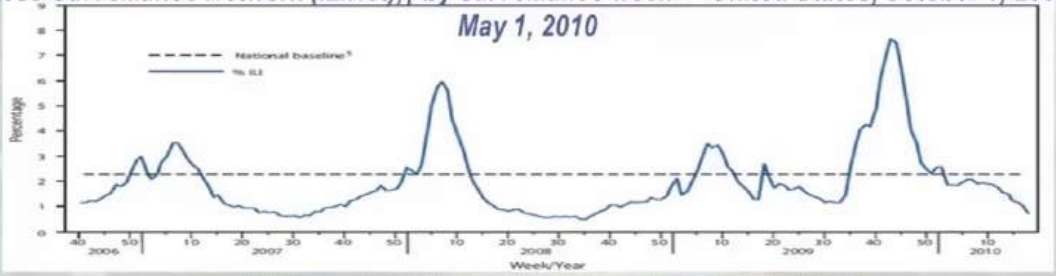


Graph by Tom

Time Series Graph

- Time series data are made of measurements for the same variable for the same individual taken at intervals over a period of time.
- Stock market prices
- Yearly rates of diseases such as influenza

Percentage of visits for influenza-like Illness (ILI) reported by the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet),† by surveillance week --- United States, October 1, 2006--May 1, 2010*



From MMWR Recommendations and Reports August 6, 2010 / 59(RR08);1-62

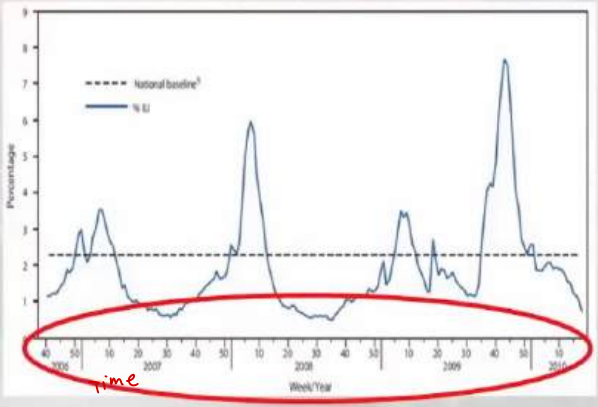
How to Plot a Time-Series Graph

1. Get time series data: one measurement per time period (month, year, etc.).

Year	Rate per 10,000
2007	25
2008	30
2009	31
2010	28
2011	20
2012	18

How to Plot a Time-Series Graph

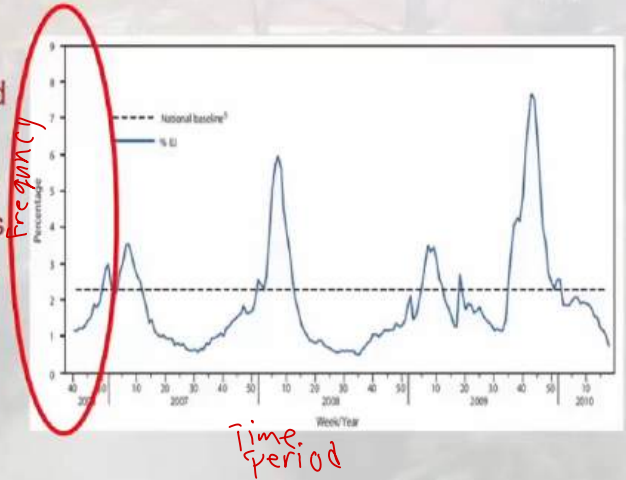
1. Get time series data: one measurement per time period (month, year, etc.).
2. Draw a horizontal line for the x-axis. Label the time periods.



From MMWR Recommendations and Reports August 6, 2010 / 59(RR08);1-62

How to Plot a Time-Series Graph

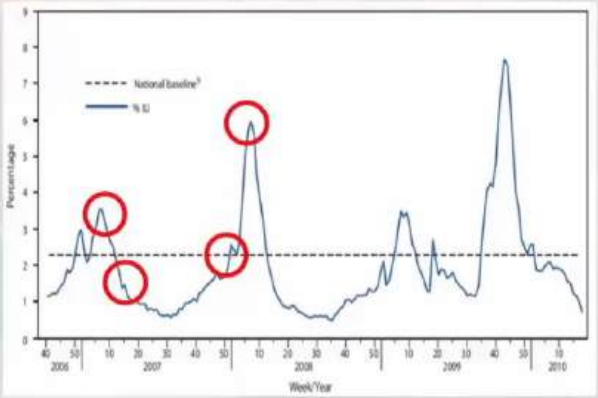
1. Get time series data: one measurement per time period (month, year, etc.).
2. Draw a horizontal line for the x-axis. Label the time periods.
3. Draw a vertical line for the y-axis. Make sure it is tall enough for the highest data value. Label it.



From MMWR Recommendations and Reports August 6, 2010 / 59(RR08);1-62

How to Plot a Time-Series Graph

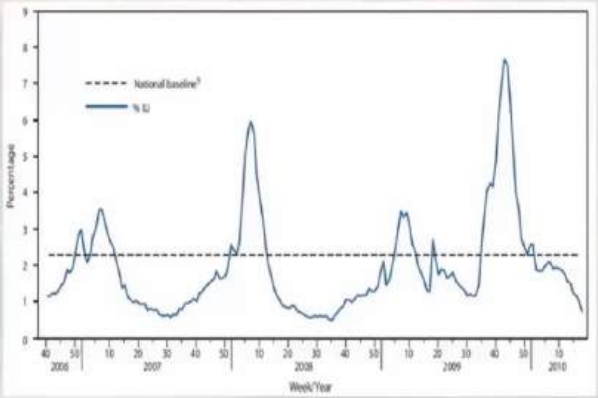
4. Looking at your data, the x-axis, and the y-axis, put dots in where the data points are.



From MMWR Recommendations and Reports August 6, 2010 / 59(RR08);1-62

How to Plot a Time-Series Graph

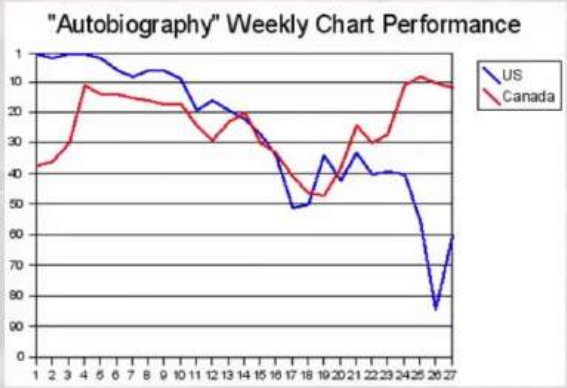
4. Looking at your data, the x-axis, and the y-axis, put dots in where the data points are.
5. Connect the dots.



From MMWR Recommendations and Reports August 6, 2010 / 59(RR08);1-62

How to Plot a Time-Series Graph

4. Looking at your data, the x-axis, and the y-axis, put dots in where the data points are.
5. Connect the dots.
6. You can have more than one line on the graph for more than one set of data values, but this requires a legend.



Graph by Camildo

Time Series Graphs

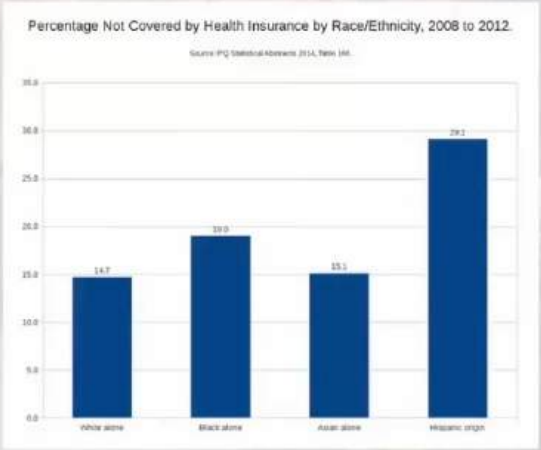
- Time series graphs are useful for understanding trends over time
- Graphing more than one set of time series data on one graph can help in comparing the differences between the data sets

Bar Graph

Display Quantitative or Qualitative Data

Bar Graph

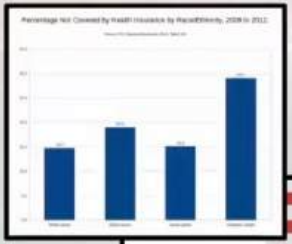
- Features of a bar graph
- How to make a bar graph
- Thinking about scale
- Pareto charts

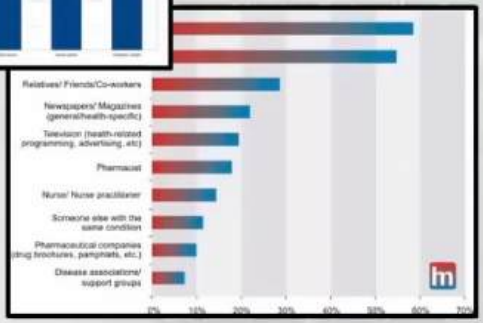


Graph by Rcragun

Features of a Bar Graph

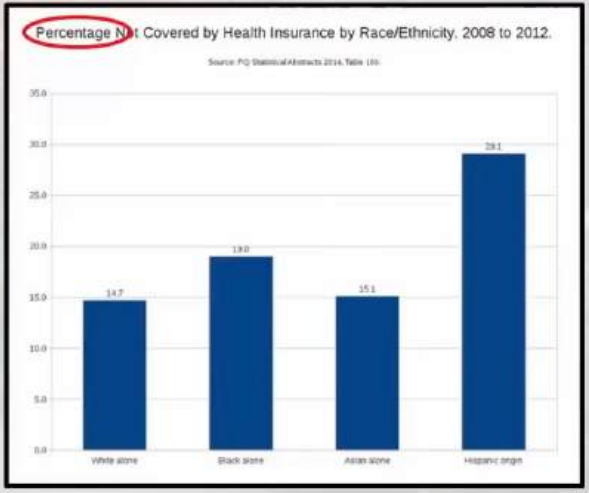
1. Bars can be vertical or horizontal
2. Are of uniform width and of uniform spacing





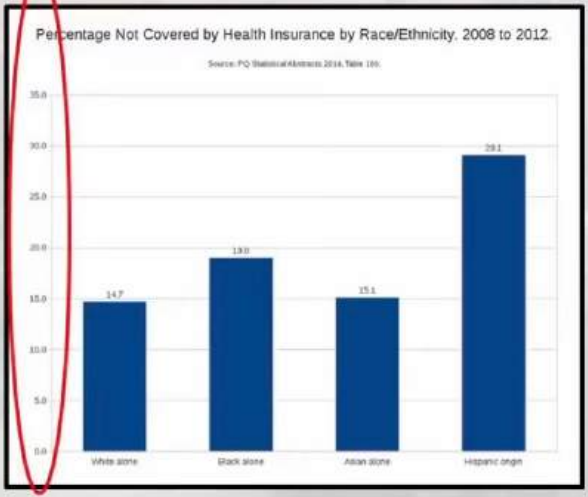
Features of a Bar Graph

1. Bars can be vertical or horizontal
2. Are of uniform width and of uniform spacing
3. Length of bars represent variable's *frequency* or *percentage* of occurrence.



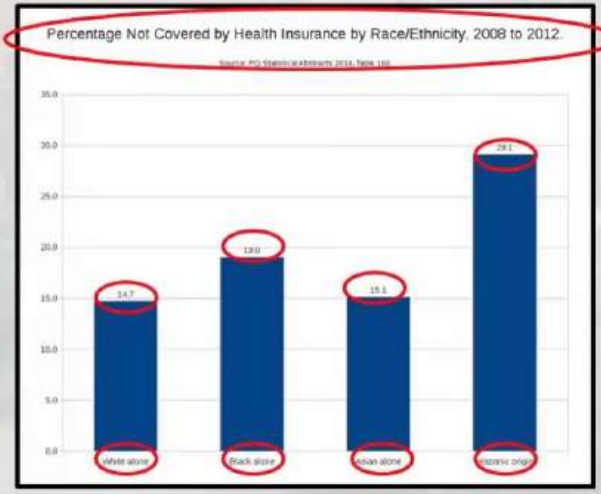
Features of a Bar Graph

1. Bars can be vertical or horizontal
2. Are of uniform width and of uniform spacing
3. Length of bars represent variable's *frequency* or *percentage* of occurrence.
4. Same measurement scale used for each bar.



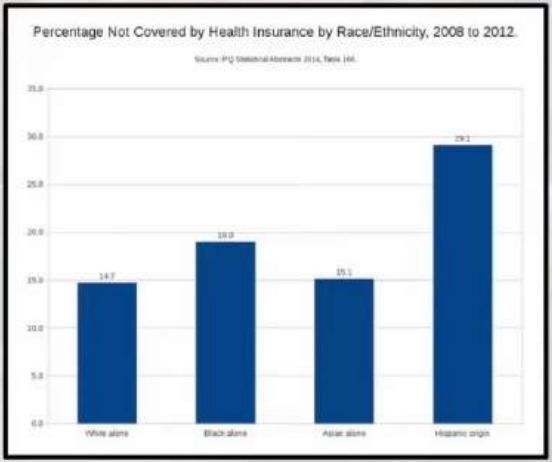
Features of a Bar Graph

1. Bars can be vertical or horizontal
2. Are of uniform width and of uniform spacing
3. Length of bars represent variable's *frequency* or *percentage* of occurrence.
4. Same measurement scale used for each bar.
5. Includes title, bar labels, and scale labels on axis or actual values for each bar.



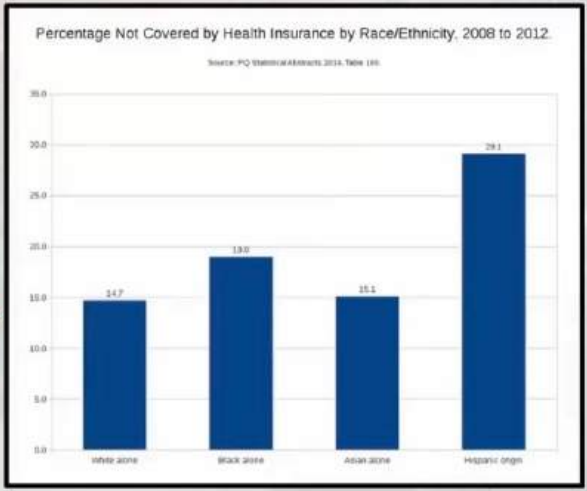
Bar Graph vs. Histogram: What's the Difference?

- The frequency histogram and relative frequency histogram are *"special cases"* of a bar graph
- They are bar graphs that:
  - *Must* have classes of a quantitative variable on the x-axis
  - *Must* have frequency or relative frequency on the y-axis

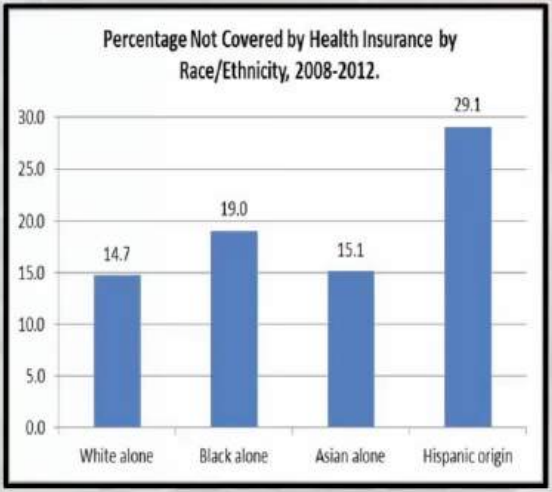


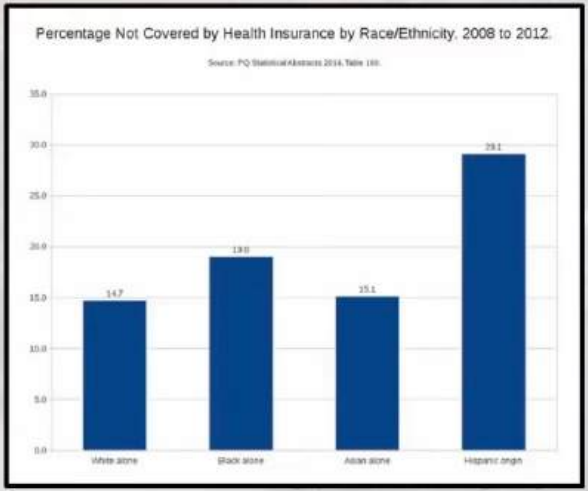
Warning About Changing the Scale

- Notice on the right side of the slide that the top number on the y-axis is 35.0.
- The highest value we have is 29.1.
- See what happens when we change the y-axis to having a maximum of 30.0 instead of 35.0.



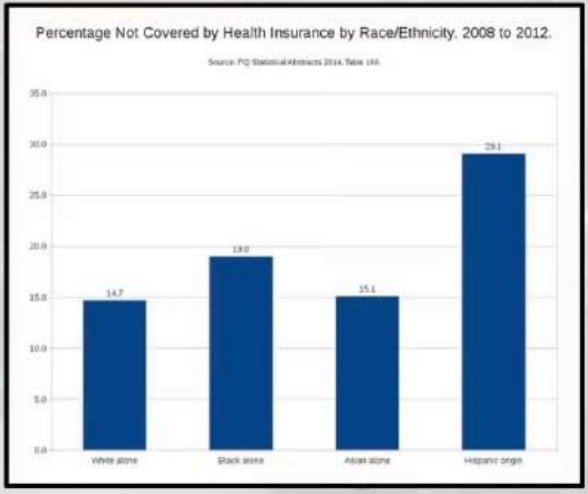
Warning About Changing the Scale





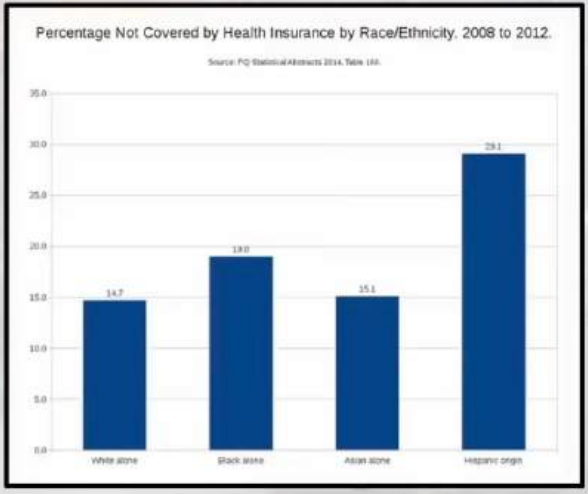
Warning About Changing the Scale

- See how the differences between bars are more dramatic when we change the scale to be shorter?
- Now, let's go the other way.
- Let's see what happens if we may the y-axis twice as high – with the maximum at 70.0.

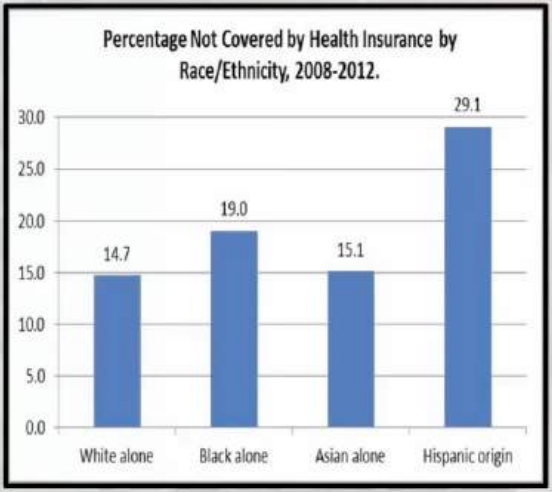


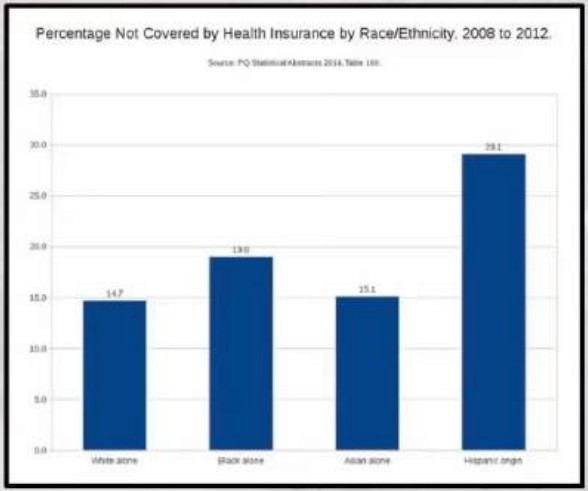
Warning About Changing the Scale

- See how the differences between bars are more dramatic when we change the scale to be shorter?
- Now, let's go the other way.
- Let's see what happens if we may the y-axis twice as high – with the maximum at 70.0.



Warning About Changing the Scale

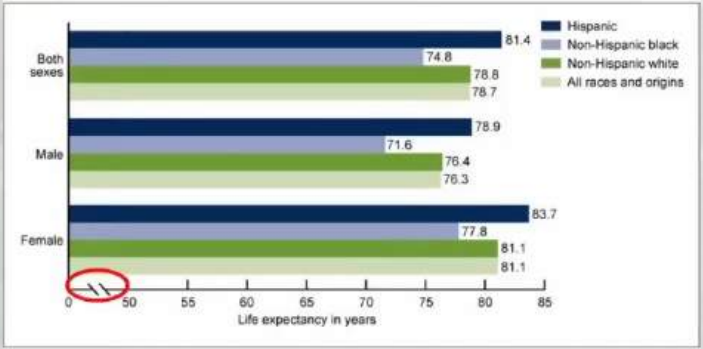




Warning About Changing the Scale

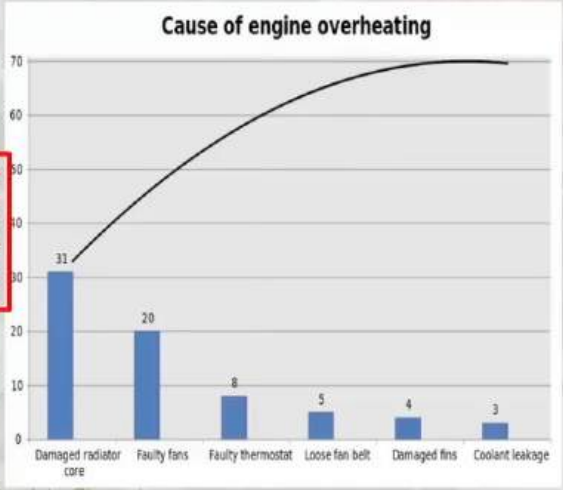
- With a taller y-axis, the differences between bars looks less dramatic.
- *Clustered* means more than one bar is graphed for each category (see legend).
- Also, look for the beginning of the scale. Some do not start at zero, and then the bars do not start at zero.
- Look for the squiggle.

Figure 1. Life expectancy at birth, by Hispanic origin, race for non-Hispanic population, and sex: United States, preliminary 2011



NOTE: Life expectancies for the Hispanic population are adjusted for underreporting of Hispanic ethnicity but are not adjusted to account for the potential effects of reverse migration.
SOURCE: National Vital Statistics System, Mortality.

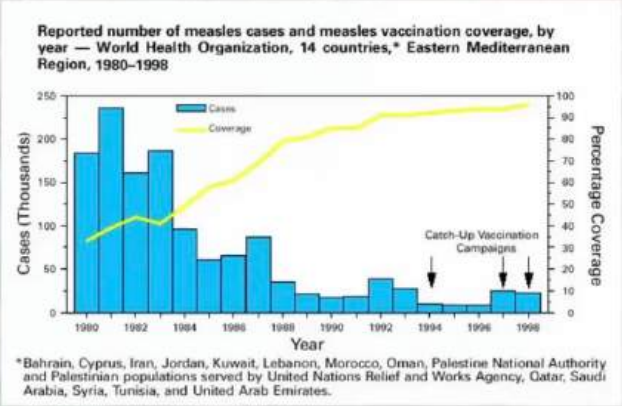
Pareto Chart



Graph by Zirguezi

- The height of the bar indicates the frequency of an event.
- Arranged left to right according to decreasing height.
- Meant to graph frequencies of “problems”
- Used more in engineering than in healthcare.

Bar Graph Summary


Reported number of measles cases and measles vaccination coverage, by year — World Health Organization, 14 countries,* Eastern Mediterranean Region, 1980–1998
Cases
Coverage
Cases (Thousands)
Percentage Coverage
Catch-Up Vaccination Campaigns
Year
*Bahrain, Cyprus, Iran, Jordan, Kuwait, Lebanon, Morocco, Oman, Palestine National Authority and Palestinian populations served by United Nations Relief and Works Agency, Qatar, Saudi Arabia, Syria, Tunisia, and United Arab Emirates.

- Bar graphs must be made following a few rules.
- Can be very helpful for visualizing and comparing quantitative and qualitative data.
- Scales are important in bar graphs!
- Pareto charts not used much in healthcare.

Pie Chart

Circular Method of Visualizing Data

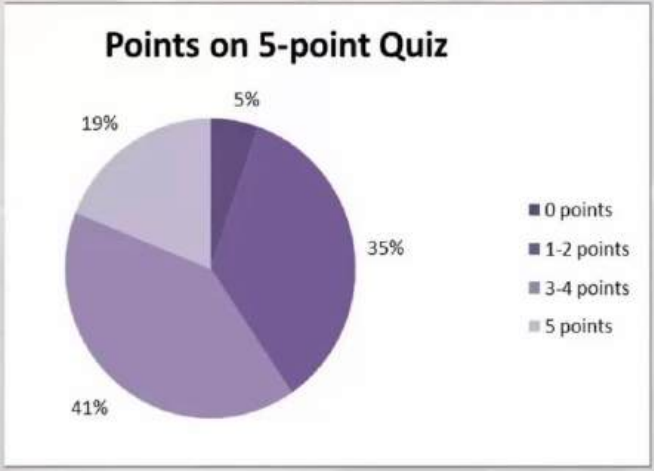
Pie Charts

- Pie chart (also circle graph) used with counts of *"mutually exclusive"* frequencies
- Often made in graphing programs because difficult to do by hand
- Very common in healthcare

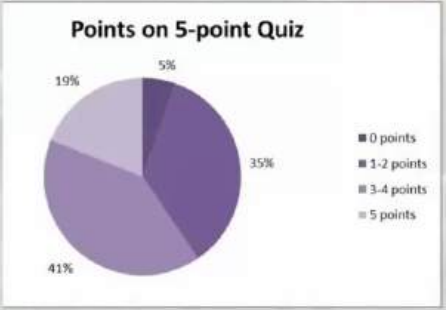
Chart from "Is a Bloodborne Pathogen Exposure Treated as an Emergency? Nurses Reveal their Experiences" By the Massachusetts Nursing Association. Available at: http://www.massnurses.org/health-and-safety/articles/bloodborne-pathogens/survey/

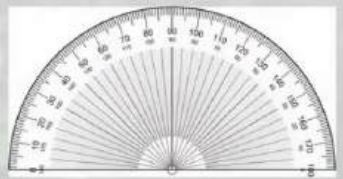
Features of a Pie Chart

- Every individual must be put in only one category.
- Can be qualitative or quantitative variable.
- If quantitative, put in classes and then graphed.



Making a Pie Chart



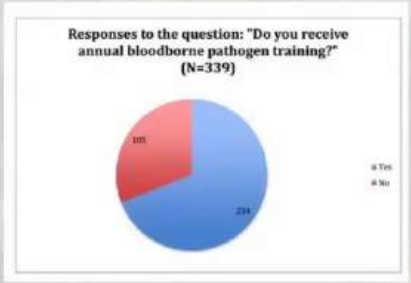


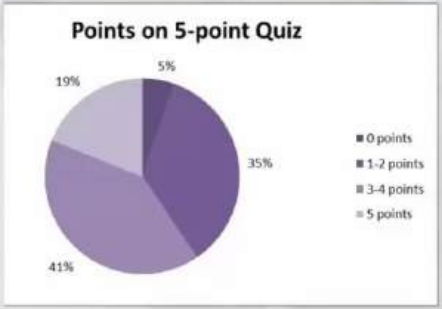
Class	Frequency	Proportion of Circle	Degrees
0 points	2	2/37=0.05	0.05*360° =19°
1-2 points	13	13/37=0.35	0.35*360° =126°
3-4 points	15	15/37=0.41	0.41*360° =146°
5 points	7	7/37=0.19	0.19*360° =68°
Total	37	1.00	360°

Protractor picture by Scientif38

A Few Notes on Pie Charts

- Must be mutually exclusive categories
  - "Favorite" color vs. "check the colors you like"
- More informative to put % than frequency, but it is helpful to do both.
- Always include title and legend.





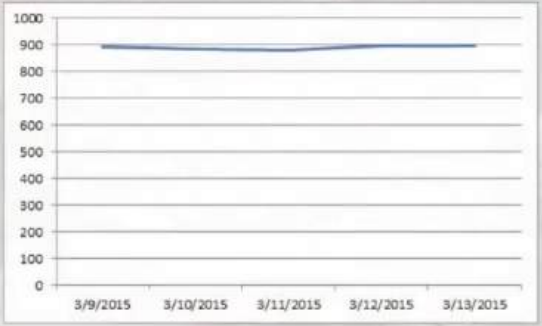
Pie Charts

Photo by Neuceu

- Pie charts are common in healthcare, and graph mutually exclusive categories.
- Easier to make using software.
- Choosing a graph requires consideration.

On All Graphs

- Provide a title
- Label axes
- Identify units of measure
- Make the graph as clear as possible (think of font size, number of items graphed).



Can you tell what is missing from the above graph?

Graph by Silverhammermba

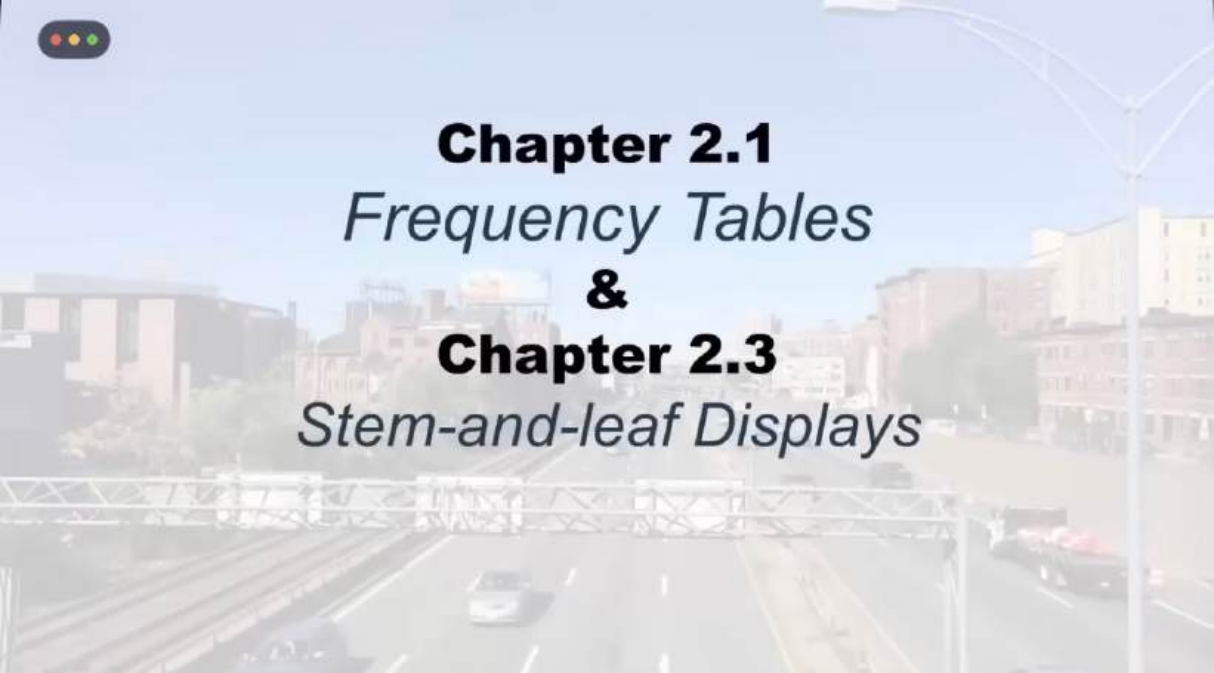
Choosing the Right Kind of Graph

Type of Graph	Cases Where Graph is Useful
Frequency Histogram	For quantitative data, when you want to see the distribution.
Relative Frequency Histogram	For quantitative data, when you want to see the distribution. Also, good for comparing to other data.
Stem-and-leaf Display	For quantitative data, when you want to see the distribution. Easier to make by hand than histogram.
Time series graph	For graphing a variable that changes over time and is measured at regular intervals.
Bar graph	For qualitative or quantitative data, and for displaying frequency or percentage.
Pareto chart	For frequencies of rare events in descending order.
Pie Graph	For mutually-exclusive categories (quantitative or qualitative).

Conclusion

- Time-series graphs show changes over time
- Bar graphs can display quantitative or qualitative data
- Pie charts are for mutually exclusive categories
- It's important to pick the right chart to provide useful visualization of the data

Pie by Evan-Amos of Vanamo Media



Chapter 2.1

Frequency Tables

&

Chapter 2.3

Stem-and-leaf Displays

Learning Objectives

t the end of this lecture, the student should be able to:

- State the steps for making a frequency table
- Define class, upper class limit, and lower class limit
- Explain what relative frequency is, and why it is useful for comparing groups
- State the steps for making a stem-and-leaf display
- Describe the difference between an “ordered” and “unordered” leaf

What is a Frequency Table?

Terms and Explanations

Frequency Tables

- Explain what a frequency table is, and why make one
- Define some terms
- Describe the steps in making a frequency and relative frequency table

Image by Tomasz Sienicki

Remember Quantitative Data?

Image by the US Navy

- Qualitative data are categorical
  - Gender, race, diagnosis
- Quantitative data are numerical
  - Age, heart rate, blood pressure

How to Organize All these Numbers?

- 60 patients were studied for the distance they needed to be transported in an ambulance.
- The shortest transport (*minimum*) was 1 mile.
- The longest transport (*maximum*) was 47 miles.
- It's hard to just look at a pile of numbers…how do we understand these data?

Photo by ibagli

A Few Definitions

- *Class:* An interval in the data.
  - Example: Between 30 and 40 miles.
- *Class limit:* The lowest and highest value that can fit in a class.
  - Example: 30 would be the *lower class limit*, and 40 would be the *upper class limit*.
- *Class width:* How wide the class is.
  - Example: Upper class limit (40) minus lower class limit (30) = 10, then add 1 = 11.
  - Example: *30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40* = 11 numbers
- *Frequency:* How many values from the data fall in the class.
  - Example: How many patients were transported 30 to 40 miles.

Decide on Classes

- Classes should be the same width
- Class width can be determined empirically
  - Example: Age 18-24, 25-34, 35-44, 45-54, 55-64, 65 and older
  - Should be based on the scientific literature

Photo by Alias 0591 from the Netherlands

Decide on Classes

- Classes should be the same width
- Class width can be determined empirically
  - Example: Age 18-24, 25-34, 35-44, 45-54, 55-64, 65 and older
  - Should be based on the scientific literature
- Can also be determined using a formula

Photo by Alias 0591 from the Netherlands

Class Width Formula

Formula

- Calculate this number: maximum – minimum.
- Divide this by the number of classes desired.
- Increase this to the next whole number

Example

- From the miles, 47 – 1 = 46.
- If we want 6 classes, 46/6 = 7.7.
- We increase this up to 8

Simple Frequency Table

- A frequency table displays each *class* along with the *frequency* (number of data points) in each class.
- Selecting arbitrary class limits can make the frequency table unbalanced.
- But not following the scientific literature can make your results non-comparable

Class Limits (Lower-Upper)	Frequency
<20 miles	41
21-29 miles	10
30-39 miles	4
40 or more miles	5
Total	60

Example for Frequency Table

Data Collection

- Glucose is measured in the blood and expressed in mg/100 ml.
- Glucose is a big molecule that should be cleared from the blood, especially if fasting.
- Blood glucose levels for a random sample of 70 women were recorded after a 12-hour fast.

Results

- Minimum = 45 mg/100 ml
- Maximum = 109 mg/100 ml
- Decided on 6 classes

Example for Frequency Table

Data Collection

- Glucose is measured in the blood and expressed in mg/100 ml.
- Glucose is a big molecule that should be cleared from the blood, especially if fasting.
- Blood glucose levels for a random sample of 70 women were recorded after a 12-hour fast.

Results

Class Limits	Frequency
45 - 55	3
56 - 66	7
67 - 77	22
78 - 88	26
89 - 99	9
100 - 110	3
Total	70

Be Careful!

- Make sure that *all* the data points are accounted for only *once* in *one* of the classes.
- Make sure the classes cover all the data.
- Make sure the total of your classes adds up to the total data points!

Photo by Erlend Schei

Relative Frequency Table

- "Relative" = in relationship to the rest of the data.
- Frequency = f
- Total sample size = n
- Relative frequency = f/n
- Relative frequency is the proportion of the values that are in that class.

Relative Frequency Table

- Relative frequency is something very useful to put in a frequency table.
- See how easy it is to calculate – take each class frequency divided by total.

Class Limits	Freq-uency	Relative Frequency
45 - 55	3	0.04
56 - 66	7	0.10
67 - 77	22	0.31
78 - 88	26	0.37
89 - 99	9	0.13
100 - 110	3	0.04
Total	70	1.00

Frequency Tables

- Frequency tables are necessary for organizing quantitative data.
- Class width must be selected, and lower and upper class limits determined
- Frequencies are then filled in.
- You can also include relative frequencies.

Photo by Robert Weißenberg

What is a Stem and Leaf?

Terms and Explanations

Stem-and-leaf

- What is a stem and leaf plot?
- How is a stem and leaf plot made?
- Why not just make a frequency table?

2	0
3	025
4	11378
5	133467889
6	024559
7	147
8	8
9	
10	2

Image by Joxemai

Why is it Called Stem and Leaf?

- In a stem and leaf, there is always a "stem"
- Leaves are then added to the stem as we tally up the length of the leaves.
- Making one will help you understand the terminology.

2	0
3	025
4	11378
5	133467889
6	024559
7	147
8	8
9	
10	2

Cornstalk photo by Huw Williams

Example: Days since Referral

- Data from 42 patients who visited a primary care clinic and were referred to mental health were collected.
- The number of days between the referral and their first mental health appointment was collected.

Substance Abuse and Mental Health Services Administration
SAMHSA
Find Help
Topics
Programs & Campaigns
Grants
Find Help
FIND YOUR CLOSEST
TREATMENT LOCATION
Behavioral Health Treatment Services Locator
Find alcohol and drug abuse treatment or mental health treatment facilities and programs around the country at findtreatment.samhsa.gov.

Building the Stem and Leaf

Days since referral

30	27	12	42	35	47
38	36	27	35	22	17
29	3	21	0	38	32
41	33	26	45	18	43
18	32	31	32	19	21
33	31	28	29	51	12
32	18	21	26	71	105

Building the Stem and Leaf

3 | 0

Days since referral

30	27	12	42	35	47
38	36	27	35	22	17
29	3	21	0	38	32
41	33	26	45	18	43
18	32	31	32	19	21
33	31	28	29	51	12
32	18	21	26	71	105

Building the Stem and Leaf

2	7
3	0

Days since referral

30	27	12	42	35	47
38	36	27	35	22	17
29	3	21	0	38	32
41	33	26	45	18	43
18	32	31	32	19	21
33	31	28	29	51	12
32	18	21	26	71	105

Building the Stem and Leaf

1	2
2	7
3	0

Days since referral

30	27	12	42	35	47
38	36	27	35	22	17
29	3	21	0	38	32
41	33	26	45	18	43
18	32	31	32	19	21
33	31	28	29	51	12
32	18	21	26	71	105

Building the Stem and Leaf

1	2
2	7
3	0
4	2

Days since referral

30	27	12	42	35	47
38	36	27	35	22	17
29	3	21	0	38	32
41	33	26	45	18	43
18	32	31	32	19	21
33	31	28	29	51	12
32	18	21	26	71	105

Building the Stem and Leaf

1	2
2	7
3	0 5
4	2

Days since referral

30	27	12	42	35	47
38	36	27	35	22	17
29	3	21	0	38	32
41	33	26	45	18	43
18	32	31	32	19	21
33	31	28	29	51	12
32	18	21	26	71	105

Building the Stem and Leaf

0	3
1	2 7
2	7 7 2 9
3	0 5 8 6 5
4	2 7

Days since referral

30	27	12	42	35	47
38	36	27	35	22	17
29	3	21	0	38	32
41	33	26	45	18	43
18	32	31	32	19	21
33	31	28	29	51	12
32	18	21	26	71	105

Building the Stem and Leaf

Later, when we get to 51, we will need to add a 5 to the stem.

0	3 0
1	2 7
2	7 7 2 9 1
3	0 5 8 6 5
4	2 7
5	1

Days since referral

30	27	12	42	35	47
38	36	27	35	22	17
29	3	21	0	38	32
41	33	26	45	18	43
18	32	31	32	19	21
33	31	28	29	51	12
32	18	21	26	71	105

Building the Stem and Leaf

Later, since there are none in the 60s, there will be a blank spot on the stem

0	3 0
1	2 7
2	7 7 2 9 1
3	0 5 8 6 5
4	2 7
5	1
6	
7	1

Days since referral

30	27	12	42	35	47
38	36	27	35	22	17
29	3	21	0	38	32
41	33	26	45	18	43
18	32	31	32	19	21
33	31	28	29	51	12
32	18	21	26	71	105

Building the Stem and Leaf

0	3 0
1	2 7
2	7 7 2 9 1
3	0 5 8 6 5
4	2 7
5	1
6	
7	1
8	
9	
10	5

At 105, the "10" is the stem.

Days since referral

30	27	12	42	35	47
38	36	27	35	22	17
29	3	21	0	38	32
41	33	26	45	18	43
18	32	31	32	19	21
33	31	28	29	51	12
32	18	21	26	71	105

Organizing Quantitative Data

Frequency Table

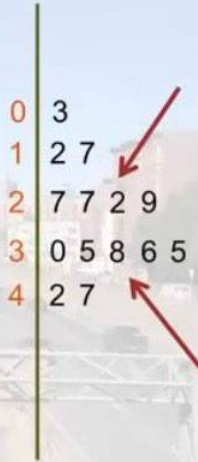
1. Need to set up classes, class widths
2. Need to count frequencies in each class
3. Lots of pre-calculations

Stem and Leaf

1. Do not need to set up classes or class widths
2. No need to count. Can tally the data as you go through the list.
3. Quicker to do

Good Idea!

- Leaf is “unordered” if numbers out of order.



Good Idea!

- Leaf is “unordered” if numbers out of order.
- After making unordered version, order the leaves.

0	3
1	2 7
2	2 7 7 9
3	0 5 5 6 8
4	2 7

Good Idea!

- Leaf is “unordered” if numbers out of order.
- After making unordered version, order the leaves.
- Then it is easier to count them up for your frequency table – no matter what classes
- Or, make each “leaf” a “class”

0	3
1	2 7
2	2 7 7 9
3	0 5 5 6 8
4	2 7

Stem-and-Leaf

- A stem and leaf is another way to organize quantitative data.
- A stem and leaf is easier to make than a frequency table and requires less preparation
- Can help you put data in order to create a frequency table

Photo by Harry Rose from South West Rocks, Australia

Conclusion

- Frequency tables and stem-and-leaf displays organize data
- Stem-and-leaf may help make a frequency table
- Purpose is to reveal "distribution" – next lecture

Painting by Paul Monnier

Chapter 3.1

Measures of Central Tendency

Learning Objectives

At the end of this lecture, the student should be able to:

- Explain how to calculate the mean.
- Describe what a mode is and say how many modes a dataset can have.
- Demonstrate how to find the median in a set of data with an odd number of values, as well as in a set of data with an even number of values.
- Define trimmed mean and weighted average.

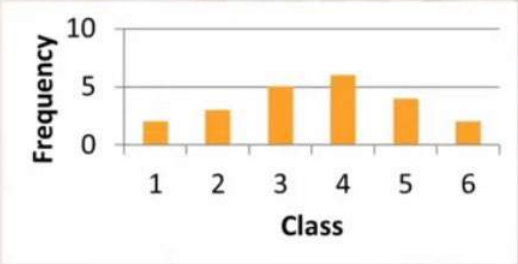
Introduction

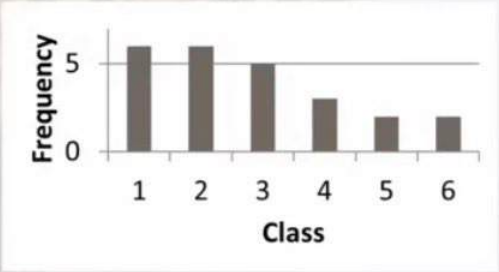
- What are “measures of central tendency”?
- Finding the mode
- Finding the median
- Calculating the mean
- Trimmed mean
- Weighted average

Image courtesy of Sarah777

Central Tendency

- How much does the data have a tendency to go to the center?
- Can only be asked about quantitative data
- Remember the distributions?





Measures of Central Tendency

- Mode
- Median
  - With odd number of values
  - With even number of values
- Mean
  - Trimmed mean
  - Weighted average

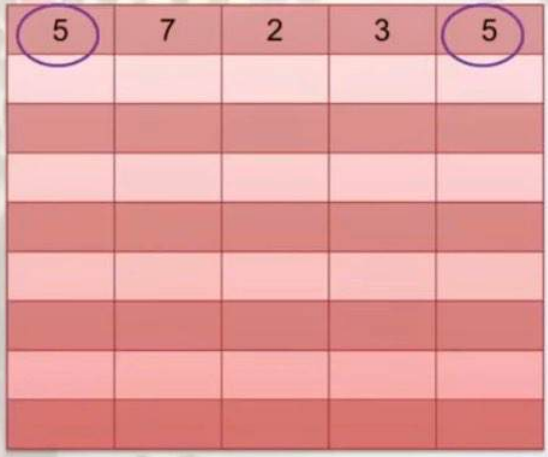
Photograph by Frank Schulenburg

Mode

Occurs Most Frequently

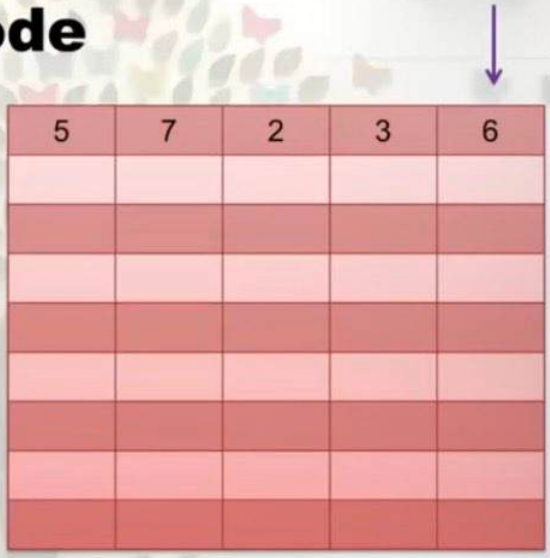
Mode

- The value that occurs most frequently in the dataset is the mode



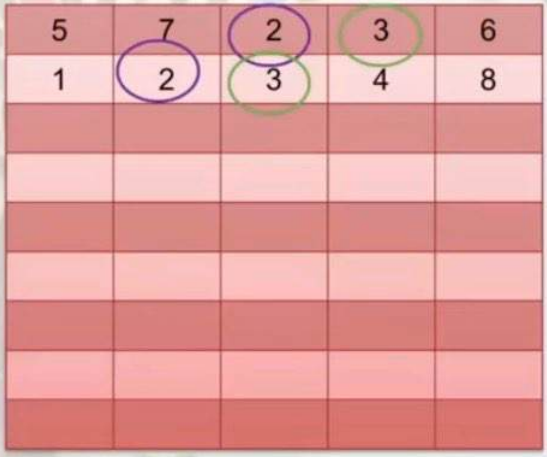
Mode

- The value that occurs most frequently in the dataset is the mode
- It is possible to have no mode.



Mode

- The value that occurs most frequently in the dataset is the mode
- It is possible to have no mode.
- It is possible to have more than one mode.



Mode

- Can get confusing with a lot of numbers in a short range – which is most numerous?
- Less confusing when scale is large
  - Systolic blood pressure
  - Platelet count

5	7	2	3	6
1	2	3	4	8
3	5	7	2	0
2	3	4	9	6
1	8	6	9	9
5	4	7	2	3
6	1	7	5	8
5	3	6	5	5
2	6	5	1	7

What Does the Mode Tell You?

- Not much
- The most "popular" answer
- The most common result
- Not used a lot in healthcare

Photograph by David Shankbone

What Does the Mode Tell You?

- Not much
- The most "popular" answer
- The most common result
- Not used a lot in healthcare

Photograph by David Shankbone

Median

Middle of the Data

Median is the Center of the Data

- Every set of quantitative data can be sorted in order of lowest to highest.
- Sometimes there are repeats in the data
- Sometimes there are outliers
- Sometimes all the data values are almost the same
- Even so, they can be arranged in order
- The “median” is the number that is halfway

How to Find the Median

1. Order the data from smallest to largest
2. If there is an odd number of values, the median is the middle data value.

1	2	3	4	5
21	33	42	62	78

Median is 42

1	2	3	4	5	6
1	3	7	8	9	12

How to Find the Median

1. Order the data from smallest to largest
2. If there is an odd number of values, the median is the middle data value.
3. If there is an even number of values, the median is the sum of the middle two values divided by 2

1	2	3	4	5
21	33	42	62	78

1	2	3	4	5	6
1	3	7	8	9	12

Median is (7+8)/2 = 7.5

How to Find the Middle Number

- After arranging the values in order, you have an "ordered set".
- If there are an odd number of values (n), take n+1, and divide by 2. Count up that many, and that is the median.

Example

Imagine n=21.

21+1 = 22, and 22/2 = 11.

You would count from the beginning of the ordered set, and the 11^{th} value in it would be the median.

How to Find the Middle Number

- After arranging the values in order, you have an "ordered set".
- If there are an even number of values (n), take n+1, and divide by 2. The number will end in something .5. Sum numbers on either side of that position and divide by 2.

Example

Imagine n=14.

14+1 = 15, and 15/2 = 7.5.

You would count from the beginning of the ordered set, and take the 7^{th} value and the 8^{th} value, add them together, then divide by 2 for the median.

What Does the Median Tell You?

- The 50th percentile of the data.
- The middle rank of the data.
- The median doesn't care much about the ends of the data. Outliers don't bother it.
- It is resistant. It is stable.

Photograph by Farragutful

Mean, Trimmed Mean & Weighted Average

Discussion of Resistant Measures

Greek Letter Capital Sigma

- Whenever you see this, say in your head, "Sum of ___"
- Σx is pronounced, "Sum of x"
- x's are the values in your data
- Means "add up all the x's"
- Another example:
  - Σxy.
  - "Sum of xy."
  - Means you must have a bunch of xy's and need to add them up.
- Σ used a lot in statistics

Sigma → sum of

Formula for Mean

- Σx = add up all the x's.
- n = number of values in your data
- After summing up all the x's, you divide by n.

$$\Sigma x/n$$

n=6

Σ (12 7 3 8 1 9) = 40

40/6 = 6.7

Notation about Means

Sample Statistics

- If your mean comes from a sample, it is called x-bar.
- Use a lowercase n for sample size

$\bar{X}$

$$\bar{X} = \frac{\Sigma x}{n}$$

Population Parameters

- If your mean comes from a population, it is called mu.
- Use an uppercase N for population size

μ μ

$$\mu = \frac{\Sigma x}{N}$$

Means vs. Medians

Medians

- Very resistant to outliers.
- Very stable.

Means

- Not resistant to outliers.
- Not very stable.

Photograph by Rob Lavinsky, iRocks.com – CC-BY-SA-3.0

Trimming the Mean

- A very high value or very low value (outliers) can really throw off the mean.
- This is not a problem with the median.
- One solution to make the mean more *resistant* is to "trim" data off each end so the outliers get cut off.
- If you trim, you have to be fair and trim the same amount off each side.

How to make a 5% Trimmed Mean

1. Figure out how many data points you have. Then, figure out what 5% of them would be.
   - If you have 100 data points, 5% would be 5 data points.
2. Put the data in order.
3. Remove 5% from the top and 5% from the bottom.
   - Remove the 5 top ones and the 5 bottom ones.
4. Now make the mean out of the remaining data.

Weighted Average

$$= \frac{\Sigma xw}{\Sigma w}$$

- Sometimes, certain values should count more toward the mean than others.
- If homework is 10% of your grade, and quizzes are 20% of your grade, the quizzes count for more than the homework
- You can arrange this by doing a weighted mean.

Example

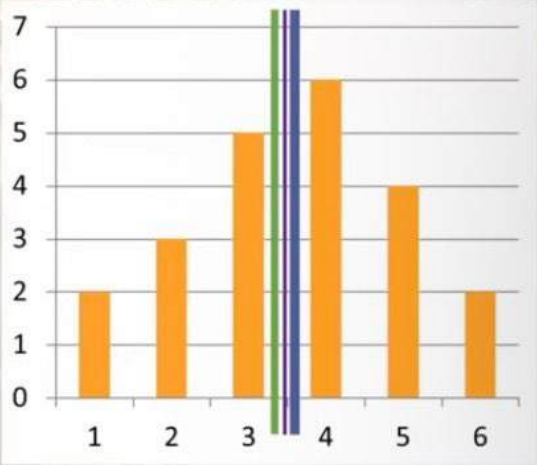
- Homework worth 10%, quizzes worth 20%, and final worth 70%.
- You got an A (4.0) on homework, B + (3.5) on quizzes, and B (3.0) on final
- Non-weighted average: (4+3.5+3)/3 = 3.5
- Weighted average: (4.0*0.1) + (3.5*0.2) + (3.0*0.7) = 3.2

$\Sigma xw = 3.2$

$\Sigma w = 1.0$

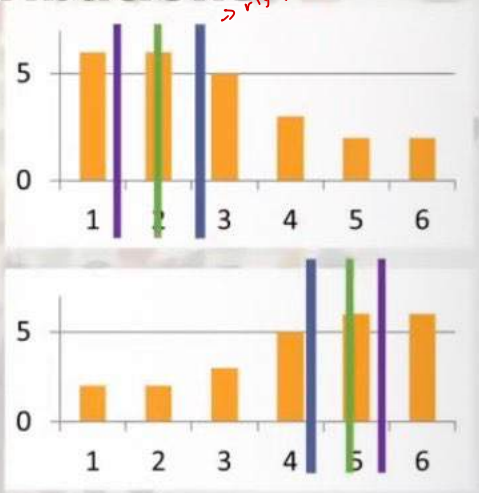
Measures of Central Tendency & Normal Distribution

- This is a normal distribution.
- Using these data, a person could determine:
  - Mean
  - Median
  - Mode
- They are all on top of each other



Measures of Central Tendency & Skewed Distributions

- Here are right and left skewed distributions.
- Note the relative positions of the following:
  - Mean
  - Median
  - Mode



Conclusion

- Measures of central tendency: Mode, median, and mean
- Mean is not resistant, median is
- Can use a trimmed mean
- Weighted average
- Distributions and measures of central tendency

Photograph by Jonathan Billinger

Chapter 3.2

Measures of Variation

Learning Objectives

At the end of this lecture, the student should be able to:

- State three different measures of variation used in statistics.
- Explain how to calculate variance and standard deviation.
- Calculate the coefficient of variation and explain its interpretation.
- State Chebyshev's Theorem.

Introduction

- Range
- Variance and standard deviation
- Coefficient of variation
- Chebyshev's Theorem
- Calculating Chebyshev Intervals

Image courtesy of Mizunoryu, Badseed, Jacobolus

Variation

- How much does the data vary?
- Imagine two different classes with the same mean grade.
- Can we tell how consistent each class was internally in terms of grades from a measure of central tendency?

Photograph by Aburk018

Measures of Variation

- Range
- Variance and Standard Deviation
  - Sample and population
- Coefficient of variation

Photograph by Glenys Gaske

Range

Simplest to Calculate

Range

- The range is the difference between the maximum and minimum value.
- It is easy to calculate
- Don't forget to actually do the subtraction

42 33 21 78 62

- In the above data, 78 is the maximum.
- 21 is the minimum.
- 78 minus 21 = 57
- 57 is the range

The Range is Not Very Useful

- The range is not very useful for looking at variation because it just relies on two points.
- It's not stable or resistant.
- Therefore, other measures are used to look at variation typically.

- We had a range of 78 minus 21 = 57
- We could just change these two numbers and get a totally different range

Variance & Standard Deviation

Defining Formula

Variance & Standard Deviation are Friends

- They are friends because this is how you calculate them:
  1. First, calculate the variance.
  2. Then, take the square root of variance, and that is the standard deviation.

Photograph by David Shankbone

- "Variance"
  - How much the data vary.
  - Think: how well does the mean represent the spread of the data?
- "Standard deviation"
  - "Standard" – following a standard, same
  - "deviation" – like a deviated septum

Formulas for Variance & Standard Deviation

- The formulas for *sample variance* and *sample standard deviation* are different than those for *population variance* and *population standard deviation*
  - We don't use the population ones that often
  - We will concentrate on how to use the sample ones
- Two different ways of doing the formula (for both sample and for population) – the "defining formula" and the "computational formula"
  - Both get the same results
  - I dislike the "computational formula" because I get really confused
  - Therefore, I will teach you the "defining formula"

Let's Look at Formulas

Sample Defining Formulas

$$\text{Sample variance} = s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

$$\text{Sample standard deviation} = s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

Photograph by Coyau

Let's Look at Formulas

Sample Defining Formulas

Sample variance $= s^2 = \dfrac{\sum(x - \bar{x})^2}{n - 1}$

Sample standard deviation $= s = \sqrt{\dfrac{\sum(x - \bar{x})^2}{n - 1}}$

- This part of the formula is called the "sum of squares"
- We will first learn how to calculate "sum of squares" so it can be entered into the formula.

Sum of Squares

Column I x	Column II x minus x-bar	Column III (x minus x-bar)^2

- $\sum(x - \overline{x})^2$ is another way of saying sum of squares.
  - An easy way to explain this (what the "squares" are) is to show an example of how to calculate it.
  - Imagine a sample of 6 patients presented to the central lab. We asked them how many minutes they had to wait to get service, and got these numbers:

2 3 3 8 10 10

Sum of Squares

Column I x	Column II x minus x-bar	Column III (x minus x-bar)^2
2	2 – 6 = -4	
3	3 – 6 = -3	
3	3 – 6 = -3	
8	8 – 6 = 2	
10	10 – 6 = 4	
10	10 – 6 = 4	
$\sum x = 36$		

- $\sum(x - \bar{x})^2$ is another way of saying sum of squares.
- To figure out Column II, the sample mean (x-bar) is needed.
- 36/6 (n=6) = a mean of 6.
- Now Column II can be filled in.

2 3 3 8 10 10

Sum of Squares

Column I x	Column II x minus x-bar	Column III (x minus x-bar)^2
2	$2 - 6 = -4$	$(-4)^2 = 16$
3	$3 - 6 = -3$	$(-3)^2 = 9$
3	$3 - 6 = -3$	$(-3)^2 = 9$
8	$8 - 6 = 2$	$(2)^2 = 4$
10	$10 - 6 = 4$	$(4)^2 = 16$
10	$10 - 6 = 4$	$(4)^2 = 16$
$\sum x = 36$		$\sum(x - \bar{x})^2 = 70$

- $\sum(x - \bar{x})^2$ is another way of saying sum of squares.
- For Column III, square each value in Column II. Those are the "squares".
- At the bottom, add them up. That's the sum of squares.
- Sum of squares = 70.

2 3 3 8 10 10

Variance Formula

$$\text{Sample variance} = s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

- Remember, the sample was from 6 patients.
- Therefore, $n = 6$
- $n - 1 = 6 - 1 = 5$.

- $\sum(x - \bar{x})^2$ is another way of saying sum of squares.
  - For Column III, square each value in Column II. Those are the "squares".
  - At the bottom, add them up. That's the sum of squares.
  - Sum of squares = 70.

2 3 3 8 10 10

Standard Deviation Formula

$\frac{70}{5} = 14$

$\sqrt{14} = 3.74$

Sample variance $= s^2 = \frac{\sum(x - \overline{x})^2}{n - 1}$

Sample standard deviation $= s = \sqrt{\frac{\sum(x - \overline{x})^2}{n - 1}}$

Let's Look at Formulas

Sample Defining Formulas

Sample variance = $s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$

Sample standard deviation = $s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$

Population Defining Formulas

Population variance = $\sigma^2 = \frac{\sum(x - \mu)^2}{N}$

Population standard deviation = $\sigma = \sqrt{\frac{\sum(x - \mu)^2}{N}}$

Coefficient of Variation

For Comparisons

Coefficient of Variation

- You will notice the word "coefficient" used a lot in statistics
- Coefficient of Variation is CV for short
- CV shows how much the data varies compared to the mean
- Always expressed in a %

$$\text{Sample CV} = \frac{s}{\bar{x}} \times 100$$

$$\text{Population CV} = \frac{\sigma}{\mu} \times 100$$

Patients:
s = 3.74, x-bar = 6

$$\frac{3.74}{6} \times 100 = 62\%$$

CV is used for Comparison

- It's hard to explain with only one group of patients.
- CV has no units – so you could compare two different ways of measuring a lab value, for example.
- The CV is the measure of the spread of the data relative to the average of the data.
  - In the first sample, the s is only 50% of the mean.
  - In the second sample, the s is 62% of the mean.

Other Patients:
s = 4, x-bar = 8

$$\frac{4}{8} \times 100 = 50\%$$

Patients:
s = 3.74, x-bar = 6

$$\frac{3.74}{6} \times 100 = 62\%$$

Chebyshev's Theorem

With Explanation of Chebyshev Intervals

What Chebyshev Figured Out

- First, he started thinking like this:
  - If you have an x-bar and an s, you can create lower and upper limits by subtracting the s and adding the s to the x-bar.
  - You can do this with a μ and σ, too – population version
- For example, if I had an μ of 100, and an σ of 5:
  - If I subtracted 1 σ from 100, I'd get 95 as the lower limit
  - If I add 1 σ to 100, I get 105 as the upper limit.
  - I could even try this by doing 2 σ – meaning subtracting 10 for the lower limit and adding 10 for the upper limit
- He realized if he used some rules along with this, there would be an interpretation of these limits that would be useful.

Chebyshev's Theorem

- He figured out the upper and lower limits, when figured out this way, explained *at least* what % of the data would be between these limits in his dataset
- He wanted this to work for all distributions, not just normal
- He used a formula to figure out this % that was based on the s. He used "k" to mean the number of "s"'s (or "σ"'s) in the equation:

$$1 - \frac{1}{k^2} = \%\text{ of data between x-bar minus k and x-bar plus k}$$

Chebyshev's Theorem

$$1 - \frac{1}{k^2} = \%\text{ of data between x-bar minus k and x-bar plus k}$$

- *Note: Chebyshev already did this, so you don't have to memorize this formula. This is just the proof.*
- He used the following numbers for k: 2, 3, and 4
- He got the following answers:
  - If you do x-bar (or μ) minus 2 s (or 2σ) and x-bar (or μ) plus 2 s (or 2σ), the % of data between the lower and upper limits is 75%.
    - If you plug 2 into the equation, you will get 75%.
  - For 3 s (or 3σ), it is 88.9%
  - For 4 s (or 4σ) it is 93.8%
- The space between the upper and lower limits is called the "Chebyshev Interval"

Let's Try Chebyshev's Intervals!

- Calculate limits for Patient Sample:
  - 75% limits: 6 +/- (2 * 3.74) = -1.48 to 13.48
  - 88.9% limits: 6 +/- (3 * 3.74) = -5.22 to 17.22
  - 93.8% limits: 6 +/- (4 * 3.74) = -8.96 to 20.96
- Interpretation
  - *At least* 75% of the data are between -1.48 and 13.48
  - *At least* 88.9% of the data are between -5.22 to 17.22
  - *At least* 93.8% of the data are between -8.96 and 20.96.

Patient Sample (waiting room minutes)
s = 3.74, x-bar = 6

s or 6	% of data in interval
2	75%
3	88.9%
4	93.8%

Let's Try Chebyshev's Intervals!

- Calculate limits for Patient Sample:
  - 75% limits: 6 +/- (2 * 3.74) = -1.48 to 13.48
  - 88.9% limits: 6 +/- (3 * 3.74) = -5.22 to 17.22
  - 93.8% limits: 6 +/- (4 * 3.74) = -8.96 to 20.96
- Interpretation
  - *At least* 75% of the data are between -1.48 and 13.48
  - *At least* 88.9% of the data are between -5.22 to 17.22
  - *At least* 93.8% of the data are between -8.96 and 20.96.

Patient Sample (waiting room minutes)
s = 3.74, x-bar = 6

s or 6	% of data in interval
2	75%
3	88.9%
4	93.8%

If dataset had 100 patients:

- *At least* 75 patients would have waited between -1.48 and 13.48 minutes
- At least 88.9 patients would have waited between -5.22 to 17.22 minutes
- At least 93.8 patients would have waited between -8.96 and 20.96 minutes.

Take-home Message on Chebyshev Interval

- It works for any distribution (normal, skewed, etc.)
- Chebyshev intervals tell you that *at least* a certain % is in the interval
- Chebyshev intervals are sometimes non-sensical (negative numbers, very high limits)
- They are not very useful and not used in healthcare
- The purpose of teaching this is to point out *in statistics, we often use the s or σ and add/subtract it from the mean because it is a good way to make lower and upper limits that have special significance.*

Conclusion

- Calculating range
- Calculating variance and standard deviation
- Calculating and using the coefficient of variation (CV)
- Chebyshev's Theorem and intervals

Stamp from USSR in Public Domain

Chapter 3.3

Percentiles and Box-and-Whisker Plots

Learning Objectives

At the end of this lecture, the student should be able to:

- Explain what a percentile means.
- Describe what the “interquartile range” is and how to calculate it.
- Explain the steps to making a box-and-whisker plot.
- State how a box-and-whisker plot helps a person evaluate the distribution of the data.

Introduction

- What are percentiles?
- What are quartiles?
- How to compute quartiles
- How to make a box-and-whisker plot

Photograph by Des Colhoun

Percentiles

Flashback! Standardized Tests

What are Percentiles?

- Quantitative data
- Remember standardized tests…
- Example: If you test at the 77th percentile, it means you did better than 77% of the people taking the test.
- If 100 people took the test, you'd have done better than 77 of them.

Photograph by KF

Percentile Definition

- Percentiles can be between 1 and 99
  - You can't have a -2nd percentile, or a 105th percentile
- Whatever number you pick:
  - That % of values fall below the number
  - And 100 minus that % of values fall above the number
- Example: 20 people take a test.
  - Let's say there is a maximum score of 5 on the test.
  - The 25th percentile means 25% of the scores fall below this score, and 75% fall above that score.
- Let's say it is an easy test, and 12 people get a 4, and the remaining 8 get a 5. The 25th percentile, or the score the cuts off the bottom 5 tests scores, will be 4. (Even the 50th percentile will be 4.)
- This would come out very different if it were a hard test, and most people got below a score of 3.

Quartiles and Interquartile Range

Specific Percentiles

Quartiles

- Quartiles is a specific set of percentiles
  - 1st quartile: 25th percentile
  - 2nd quartile: 50th percentile (also median!)
  - 3rd quartile: 75th percentile
- These can be calculated by hand.

Image from the US Mint

Computing Quartiles

1. Order the data from smallest to largest.
2. Find the median.
   - 2nd quartile
   - 50th percentile
3. Find the median of the lower half of the data.
   - 1st quartile
   - 25th percentile
4. Find the median of the upper half of the data.
   - 3rd quartile
   - 75th percentile

- Remember the *range*?
- **New!** *Interquartile range*
- Once you have 3rd quartile and 1st quartile you can calculate interquartile range (IQR)
- 3rd quartile minus 1st quartile = IQR

Quartile Example

1. Order the data from smallest to largest.

The following are a *sample* of numbers of beds from 11 Mass. hospitals (ordered).
From ahd.com

41 74 90 97 121
126 142 155 254 318
364

Quartile Example

1. Order the data from smallest to largest.
2. Find the median. That's Quartile 2/50th percentile.

The following are a *sample* of numbers of beds from 11 Mass. hospitals (ordered).

From ahd.com

41 74 90 97 121
126 142 155 254 318
364

126 = 50th percentile

Quartile Example

1. Order the data from smallest to largest.
2. Find the median. That's Quartile 2/50th percentile.

The following are a *sample* of numbers of beds from 11 Mass. hospitals (ordered).

From ahd.com

41 74 90 97 121
126 142 155 254 318
364

126 = 50th percentile

Quartile Example

1. Order the data from smallest to largest.
2. Find the median. That's Quartile 2/50th percentile.
3. Find the median of the lower half of the data for Quartile 1/25th percentile.

The following are a *sample* of numbers of beds from 11 Mass. hospitals (ordered).
From ahd.com

41 74 90 97 121
126 142 155 254 318
364

126 = 50th percentile
90 = 25th percentile

Quartile Example

1. Order the data from smallest to largest.
2. Find the median. That's Quartile 2/50th percentile.
3. Find the median of the lower half of the data for Quartile 1/25th percentile.
4. Find the median of the upper half of the data for Quartile 3/75th percentile.

The following are a *sample* of numbers of beds from 11 Mass. hospitals (ordered).
From ahd.com

41 74 90 97 121
126 142 155 254 318
364

126 = 50th percentile
90 = 25th percentile
254 = 75th percentile

Quartile Example

1. Order the data from smallest to largest.
2. Find the median. That's Quartile 2/50th percentile.
3. Find the median of the lower half of the data for Quartile 1/25th percentile.
4. Find the median of the upper half of the data for Quartile 3/75th percentile.

IQR = 254 – 90 = 164

The following are a *sample* of numbers of beds from 11 Mass. hospitals (ordered).
From ahd.com

41 74 90 97 121
126 142 155 254 318
364

126 = 50th percentile
90 = 25th percentile
254 = 75th percentile

Quartile Example

1. Order the data from smallest to largest.
2. Find the median. That's Quartile 2/50th percentile.
3. Find the median of the lower half of the data for Quartile 1/25th percentile.
4. Find the median of the upper half of the data for Quartile 3/75th percentile.

IQR = 254 – 90 = 164

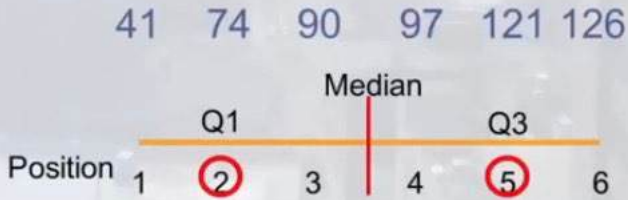
The following are a *sample* of numbers of beds from 11 Mass. hospitals (ordered).
From ahd.com

41 74 90 97 121
126 142 155 254 318
364

126 = 50th percentile
90 = 25th percentile
254 = 75th percentile

More Notes on Q1 and Q3

- Imagine we started with 6 values.
  - The median would be between the 3rd and 4th position.
  - Therefore, all 3 values below the median would be considered in calculating Q1, and all 3 above the median would be considered in calculating Q3.

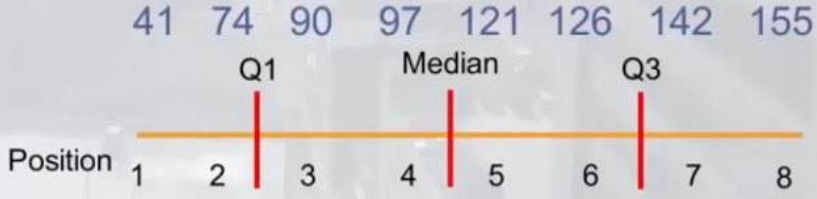


More Notes on Q1 and Q3

- What if we had 7 values instead of 6?
  - The median would be the 4th one in order. Because 7 is odd, the median then is not included in the lower or upper half of the data (used to calculate Q1 and Q3).
  - Q1 would consider the 3 values below the 4th position, and Q3 would consider the 3 values above the 4th position, but the actual median in the 4th position (median) would not be considered.

More notes on Q1 and Q3

- What if we had 8 values instead of 7?
  - The median would be the 4th and 5th positions added together and divided by 2.
  - Q1 would consider the 4 values below median, and Q3 would consider the 4 values above the median. In this case, for Q1, because there are 4 values in the lower half, the values in the 2nd and 3rd positions must be added together and divided by 2. Similarly, for Q3, this must be done for the values in the 6th and 7th positions.



More notes on Q1 and Q3

- What if we had 9 values instead of 8?
  - The median would be in the 5th position.
  - This would leave 4 values below the median, so Q1 would be between the 2nd and 3rd value, and Q3 would be between the 7th and 8th value.

Box Plot Ingredients

Units: Hospital Beds

Minimum = 41

Q1 = 90

Median (Q2) = 126

Q3 = 254

Maximum = 364

Let's make a box plot!

The following are a sample of numbers of beds from 11 Mass. hospitals (ordered)

41 74 90 97 121
126 142 155 254 318
364

Box Plot Ingredients

Units: Hospital Beds

Minimum = 41

Q1 = 90

Median (Q2) = 126

Q3 = 254

Maximum = 364

Box Plot Ingredients

Units: Hospital Beds

Minimum = 41

Q1 = 90

Median (Q2) = 126

Q3 = 254

Maximum = 364

Box Plot Ingredients

Units: Hospital Beds

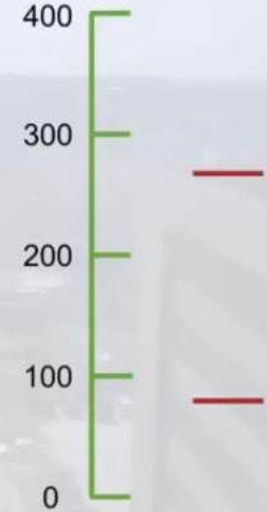
Minimum = 41

Q1 = 90

Median (Q2) = 126

Q3 = 254

Maximum = 364



Box Plot Ingredients

Units: Hospital Beds

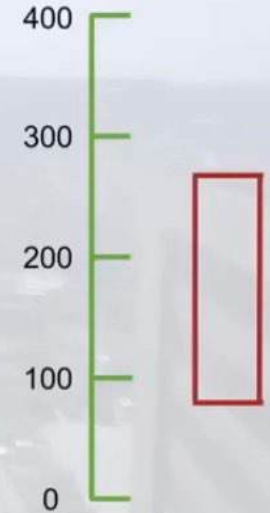
Minimum = 41

Q1 = 90

Median (Q2) = 126

Q3 = 254

Maximum = 364



Box Plot Ingredients

Units: Hospital Beds

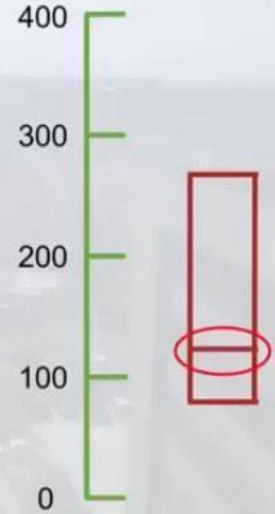
Minimum = 41

Q1 = 90

Median (Q2) = 126

Q3 = 254

Maximum = 364



Box Plot Ingredients

Units: Hospital Beds

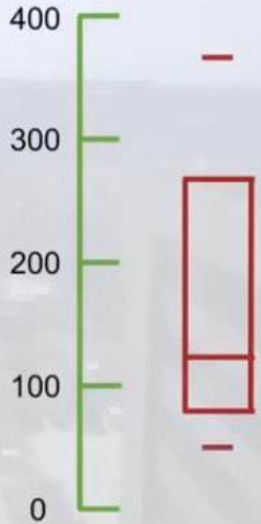
Minimum = 41

Q1 = 90

Median (Q2) = 126

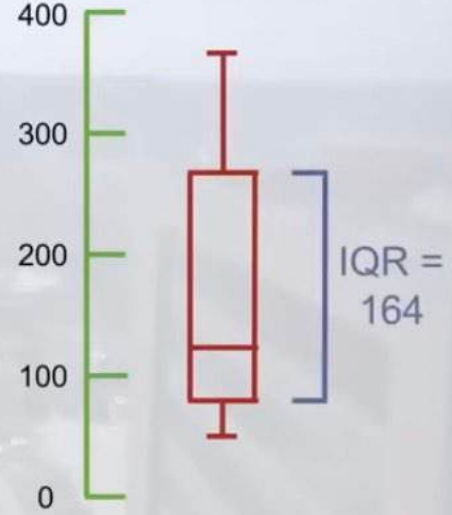
Q3 = 254

Maximum = 364



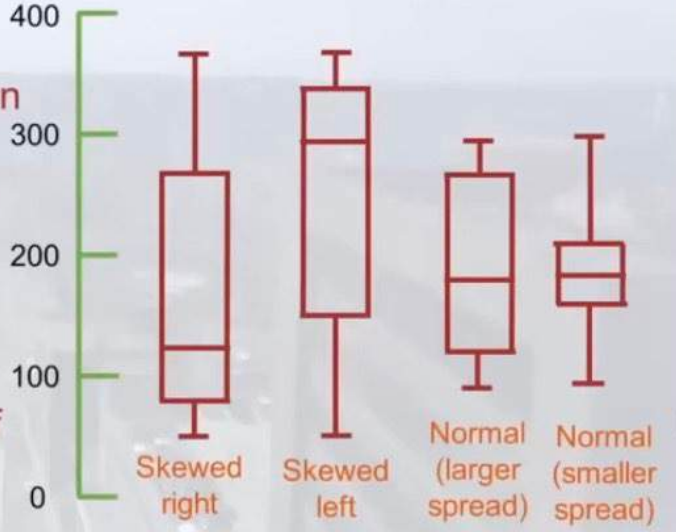
Box Plot Ingredients

Units: Hospital Beds
Minimum = 41
Q1 = 90
Median (Q2) = 126
Q3 = 254
Maximum = 364



Why Box-and-Whisker Plots?

- Help see the distribution in the data
- Another way to look at distribution
  - Histogram
  - Stem-and-leaf
- Also can see spread of data



Conclusion

- Discussion of percentiles
- Quartiles are a specific set of percentiles
- How to calculate the quartiles
- Interquartile range
- Making and interpreting a box-and-whisker plot

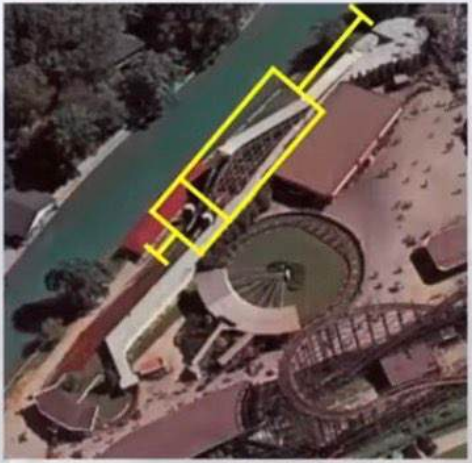


Photo courtesy of Hershey Information Center

Chapter 4.1

Scatter Diagrams and Linear Correlation

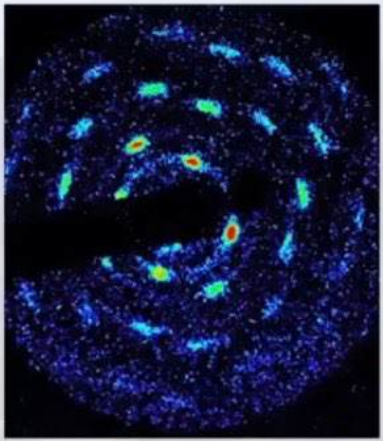
Learning Objectives

At the end of this lecture, the student should be able to:

- Explain what a scattergram is and how to make one
- State what "strength" and "direction" mean with respect to correlations
- Compute correlation coefficient r using the computational formula
- Describe why correlation is not necessarily causation

Introduction

- Making a scatter diagram
- Correlation coefficient r
- Causation and lurking variables



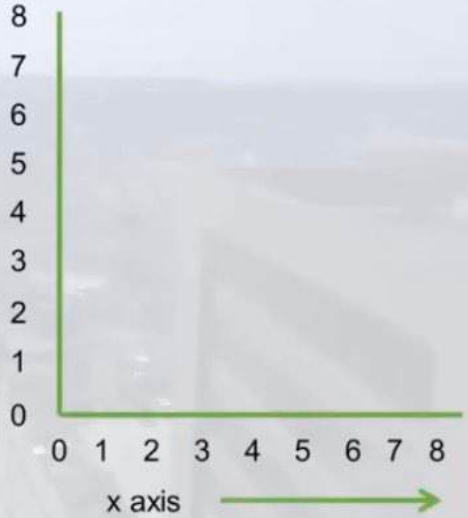
Photograph provided by Dr. John Bollinger

Scattergram

Also called Scatter Plots

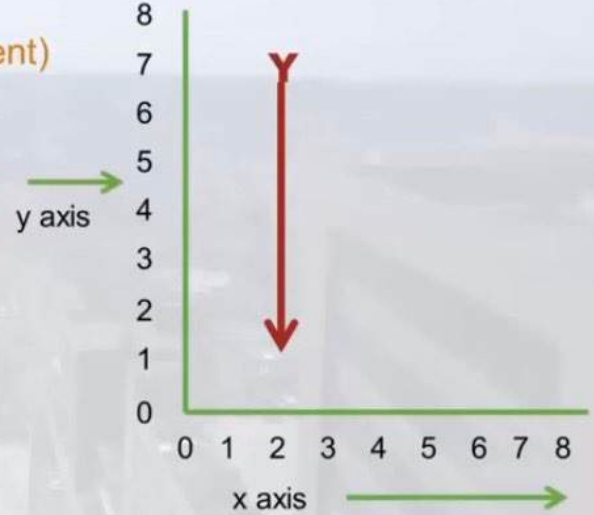
Scattergrams Graph x,y Pairs

- Explanatory (independent) variable is called x
  - Graphed on x-axis



Scattergrams Graph x,y Pairs

- Explanatory (independent) variable is called x
  - Graphed on x-axis
- Response (dependent) variable is called y
  - Graphed on y-axis



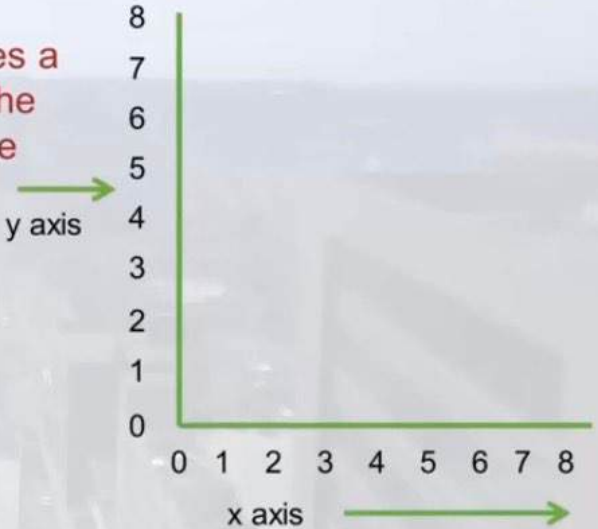
Scattergrams Graph x,y Pairs

- Explanatory (independent) variable is called x
  - Graphed on x-axis
- Response (dependent) variable is called y
  - Graphed on y-axis
- Trick to memorizing: x → y, x comes before y, so x "causes" y.
- Scatter diagram is a graph of these x,y pairs

Scattergrams Graph x,y Pairs

Do the number of diagnoses a patient has correlate with the number of medications s/he takes?

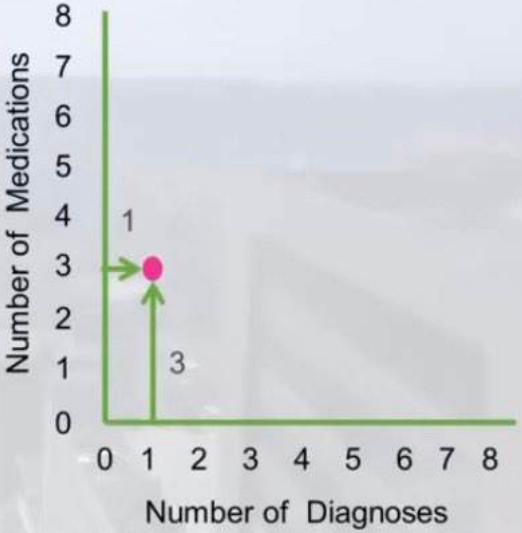
x (# of dx)	y (# of meds)
1	3
3	5
4	4
7	6



Scattergrams Graph x,y Pairs

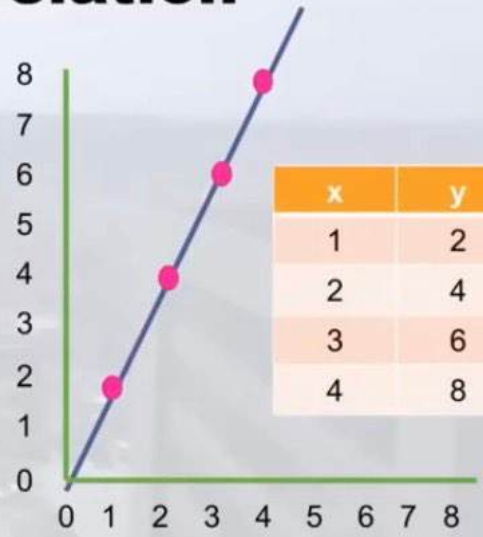
Do the number of diagnoses a patient has correlate with the number of medications s/he takes?

x (# of dx)	y (# of meds)
1	3
3	5
4	4
7	6



Linear Correlation

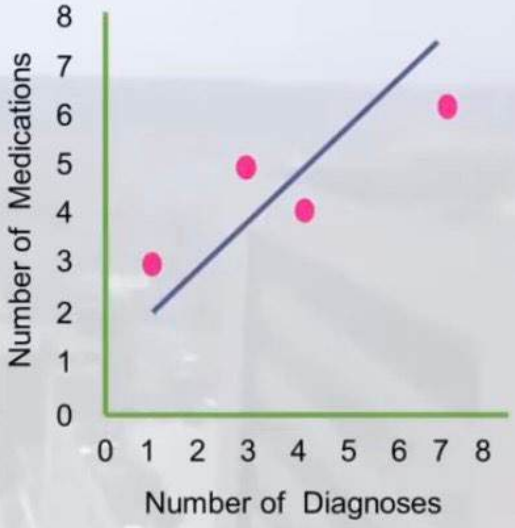
- Linear correlation means that when you make a scatterplot of x,y pairs, it looks kind of like a line
- "Perfect" linear correlation looks like graphing points in algebra



x	y
1	2
2	4
3	6
4	8

Facts About Linear Correlation

- The line can go up. This is a positive correlation.

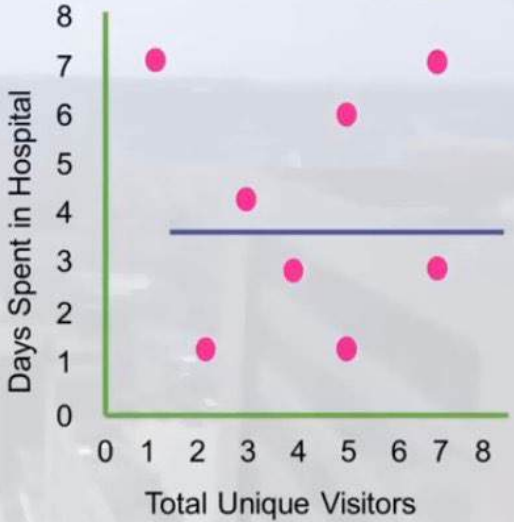


Facts About Linear Correlation

- The line can go up. This is a positive correlation.
- The line can go down. This is negative correlation.

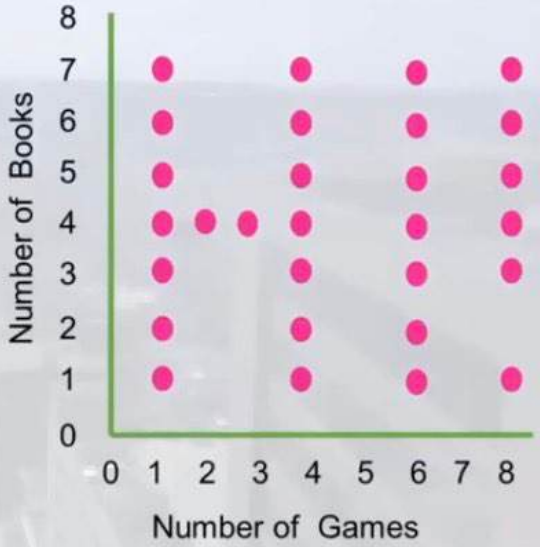
Facts About Linear Correlation

- The line can go up. This is a positive correlation.
- The line can go down. This is negative correlation.
- The line can be straight. This is no correlation.



Facts About Linear Correlation

- The line can go up. This is a positive correlation.
- The line can go down. This is negative correlation.
- The line can be straight. This is no correlation.
- The line can be goofy. This is also no correlation.



Correlation Has Two Attributes

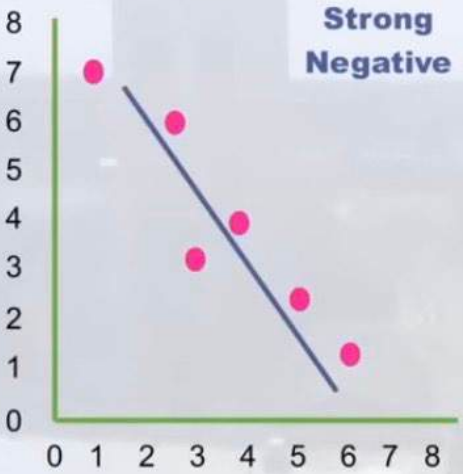
Direction

- Positive correlation
- Negative correlation
- No correlation

Strength

- Strength refers to how close to the line all the dots fall.
- If they fall really close to the line, it is strong
- If they fall kind of close to the line, it is moderate
- If they aren't very close to the line, it is weak

Correlation Has Two Attributes



Strength

- Strength refers to how close to the line all the dots fall.
- If they fall really close to the line, it is strong
- If they fall kind of close to the line, it is moderate
- If they aren't very close to the line, it is weak

Correlation Has Two Attributes

Strength

- Strength refers to how close to the line all the dots fall.
- If they fall really close to the line, it is strong
- If they fall kind of close to the line, it is moderate
- If they aren't very close to the line, it is weak

Correlation Has Two Attributes

Hey, what's that?? Outlier!

Strength

- Strength refers to how close to the line all the dots fall.
- If they fall really close to the line, it is strong
- If they fall kind of close to the line, it is moderate
- If they aren't very close to the line, it is weak

Outliers in Correlation

- Outliers can have a very powerful effect on a correlation
- An outlier in any of the 4 corners of the plot can really affect the direction of the line
- An outlier can also change the correlation from strong and moderate to weak
- It's good to look at a scatterplot to make sure you identify outliers

Correlation Coefficient r

Putting a Number on Correlation

Correlation Coefficient r

- Remember “coefficient” from CV (coefficient of variation)?
- Coefficient just means a number
- r stands for the sample correlation coefficient
  - Remember! Corrrrrrrrrrrrrrrrrrelation
  - Population correlation coefficient = ρ
- We will only focus on r

What is r?

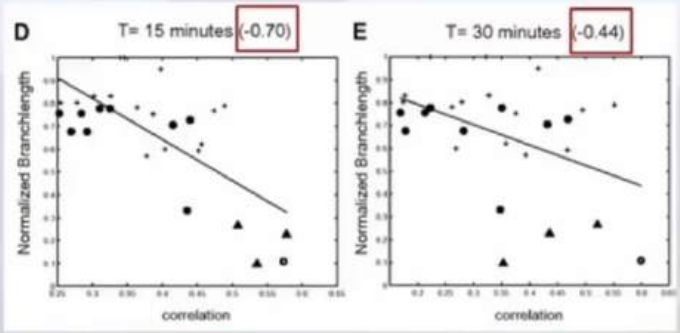
What it is

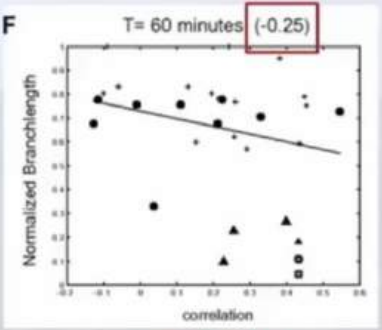
- A numerical quantification of how correlated a set of x,y pairs are
- Calculated from plugging x,y pairs into an equation
- Has a defining formula and a computational formula
- I will demonstrate computational formula

How to interpret it

- The r calculation produces a number
- The lowest number possible is -1.0
  - Perfect negative correlation
- The highest possible number is 1.0
  - Perfect positive correlation
- All others are in-between

Examples of Negative r





r = -0.70

r = -0.44

r = -0.25

OPINION!!!

For negative correlations:

- 0.0 to -0.40: Weak
- -0.40 to -0.70: Moderate
- -0.70 to -1.0: Strong

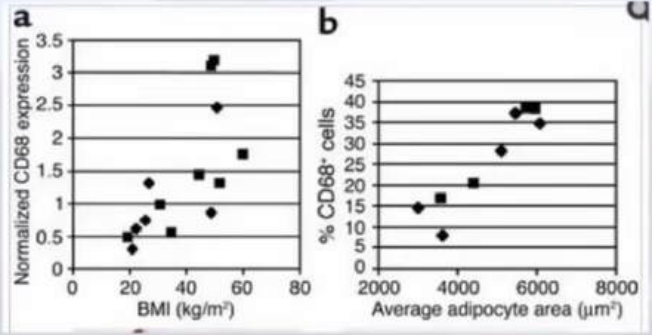
Evolutionary principles of modular gene regulation in yeasts

Figure 11.

Dawn A Thompson, Sushmita Roy, Michelle Chan, Mark P Styczynsky, Jenna Pfiffner, Courtney French, Amanda Socha, Anna Thielke, Sara Napolitano, Paul Muller, Manolis Kellis, Jay H Konieczka, Ilan Wapinski, Aviv Regev

Broad Institute of MIT and Harvard, United States; Massachusetts Institute of Technology, United States; Howard Hughes Medical Institute, Massachusetts Institute of Technology, United States

Examples of Positive r



r = 0.66 **r = 0.92**

Obesity is associated with macrophage accumulation in adipose tissue

Stuart P. Weisberg,[1] Daniel McCann,[1] Manisha Desai,[2] Michael Rosenbaum,[1] Rudolph L. Leibel,[1,3,4] and Anthony W. Ferrante, Jr.[3,4]

OPINION!!!

For positive correlations:

- 0.0 to 0.40: Weak
- 0.40 to 0.70: Moderate
- 0.70 to 1.0: Strong

Calculating r

Computational Formula

Computational Formula

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\ \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

Hypothetical Scenario

- We have 7 patients
- They have come to the clinic for appointments throughout the year.
- We predict those with a higher diastolic blood pressure (DBP) will have more appointments
- We take DBP at last appointment as "x"
- We take number of appointments over the year as "y"

- *FLASHBACK!* …to Chapter 3.2
- Notice all the Σ's
- As before, we will
  - make columns
  - make calculations
  - Then add up the columns to get these Σ's

x=DBP, y=# of Appointments

#	x	y	x^2	y^2	xy
1	70	3			
2	115	45			
3	105	21			
4	82	7			
5	93	16			
6	125	62			
7	88	12			
	$\Sigma x =$ 678	$\Sigma y =$ 166			

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\ \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

x=DBP, y=# of Appointments

#	x	y	x^2	y^2	xy
1	70	3			
2	115	45			
3	105	21			
4	82	7			
5	93	16			
6	125	62			
7	88	12			
	$\Sigma x =$ 678	$\Sigma y =$ 166			

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

x=DBP, y=# of Appointments

#	x	y	x^2	y^2	xy
1	70	3			
2	115	45			
3	105	21			
4	82	7			
5	93	16			
6	125	62			
7	88	12			
	$\Sigma x =$ 678	$\Sigma y =$ 166			

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

NOT!

Σxy will go here

x=DBP, y=# of Appointments

#	x	y	x^2	y^2	xy
1	70	3			
2	115	45			
3	105	21			
4	82	7			
5	93	16			
6	125	62			
7	88	12			
	$\Sigma x =$ 678	$\Sigma y =$ 166			

Σx^2 will go here

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

How to remember the difference between Σx^2 and $(\Sigma x)^2$:

- Do what's in () first
- So, if you got $(\Sigma x)^2$, you know what to do – take $\Sigma x * \Sigma x$
- But what if you have no ()?
  - Then you have Σx^2
  - Tell yourself it's NOT $\Sigma x * \Sigma x$ then because that would be the one with ()
  - Therefore, it must be the Σ of the x^2 column.

x=DBP, y=# of Appointments

#	x	y	x^2	y^2	xy
1	70	3			
2	115	45			
3	105	21			
4	82	7			
5	93	16			
6	125	62			
7	88	12			
	$\Sigma x =$ 678	$\Sigma y =$ 166			

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

x=DBP, y=# of Appointments

#	x	y	x^2	y^2	xy
1	70	3	4,900		
2	115	45	13,225		
3	105	21	11,025		
4	82	7	6,724		
5	93	16	8,649		
6	125	62	15,625		
7	88	12	7,744		
	Σx = 678	Σy = 166	Σx^2 = 67,892		

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\ \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

x=DBP, y=# of Appointments

#	x	y	x^2	y^2	xy
1	70	3	4,900	9	
2	115	45	13,225	2,025	
3	105	21	11,025	441	
4	82	7	6,724	49	
5	93	16	8,649	256	
6	125	62	15,625	3,844	
7	88	12	7,744	144	
	$\Sigma x =$ 678	$\Sigma y =$ 166	$\Sigma x^2 =$ 67,892	$\Sigma y^2 =$ 6,768	

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\ \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

x=DBP, y=# of Appointments

#	x	y	x^2	y^2	xy
1	70	3	4,900	9	210
2	115	45	13,225	2,025	5,175
3	105	21	11,025	441	2,205
4	82	7	6,724	49	574
5	93	16	8,649	256	1,488
6	125	62	15,625	3,844	7,750
7	88	12	7,744	144	1,056
	$\Sigma x =$ 678	$\Sigma y =$ 166	$\Sigma x^2 =$ 67,892	$\Sigma y^2 =$ 6,768	$\Sigma xy =$ 18,458

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\ \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

x=DBP, y=# of Appointments

#	x	y	x^2	y^2	xy
1	70	3	4,900	9	210
2	115	45	13,225	2,025	5,175
3	105	21	11,025	441	2,205
4	82	7	6,724	49	574
5	93	16	8,649	256	1,488
6	125	62	15,625	3,844	7,750
7	88	12	7,744	144	1,056
	$\Sigma x =$ 678	$\Sigma y =$ 166	$\Sigma x^2 =$ 67,892	$\Sigma y^2 =$ 6,768	$\Sigma xy =$ 18,458

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\ \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$\frac{(7)(18,458) - (678)(166)}{\sqrt{(7)(67,892) - (678)^2}\sqrt{(7)(6,768) - (166)^2}}$$

x=DBP, y=# of Appointments

#	x	y	x^2	y^2	xy
1	70	3	4,900	9	210
2	115	45	13,225	2,025	5,175
3	105	21	11,025	441	2,205
4	82	7	6,724	49	574
5	93	16	8,649	256	1,488
6	125	62	15,625	3,844	7,750
7	88	12	7,744	144	1,056
	$\Sigma x =$ 678	$\Sigma y =$ 166	$\Sigma x^2 =$ 67,892	$\Sigma y^2 =$ 6,768	$\Sigma xy =$ 18,458

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$\frac{(7)(18,458) - (678)(166)}{\sqrt{(7)(67,892) - (678)^2}\sqrt{(7)(6,768) - (166)^2}}$$

$$\frac{16,658}{124.74 * 140.78} = 17,561.3$$

x=DBP, y=# of Appointments

#	x	y	x^2	y^2	xy
1	70	3	4,900	9	210
2	115	45	13,225	2,025	5,175
3	105	21	11,025	441	2,205
4	82	7	6,724	49	574
5	93	16	8,649	256	1,488
6	125	62	15,625	3,844	7,750
7	88	12	7,744	144	1,056
	$\Sigma x =$ 678	$\Sigma y =$ 166	$\Sigma x^2 =$ 67,892	$\Sigma y^2 =$ 6,768	$\Sigma xy =$ 18,458

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\ \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$\frac{(7)(18,458) - (678)(166)}{\sqrt{(7)(67,892) - (678)^2}\sqrt{(7)(6,768) - (166)^2}}$$

$$\frac{16,658}{124.74 * 140.78 = 17,561.3}$$

$$r = \frac{16,658}{17,561.3} = 0.949$$

OPINION! 0.70 to 1.0: Strong

Facts About r

- r requires data with a "bivariate normal distribution" – we do not cover looking at this in this class, but please know this.
- r does not have units.
- Perfect linear correlation is r=-1.0 or r=1.0 (depending on direction). No linear correlation is r=0.
- Positive r means as x goes up, y goes up, and as x goes down, y goes down.
- Negative r means as x goes up, y goes down, and as x goes down, y goes up.
- Even if you switched x and y on the axes, you'd get the same r.
- Even if you converted x and y to different units (e.g., you converted measurements into the metric system), you'd get the same r.

Lurking Variables and "Correlation is not Causation"

Don't be Misled by Correlations!

Correlation is not Causation

- Beware of lurking variables!
- Selecting x and y is political – you are implying x could cause y
  - Example: Taller people are heavier, so x=height and y=weight
  - People who are overweight do not suddenly grow taller
- But there are other causes of weight besides height.
- Genetics can cause both height and weight.
  - A genetic profile that leads to tallness and obesity could be a lurking variable in the relationship between height and weight.

Examples

Claim

- Eating ice cream causes murders, because when more ice cream is sold, murder rates rise.

Reality

- "Summer" and warm weather are lurking variables.
  - Summer increases ice cream consumption
  - Summer means more people are outside so more murders occur.

Examples

Claim

- Over time, as people purchase more onions, the stock market rises. This is true for many generations in the US.

Reality

- "A healthy economy" is the lurking variable
  - A healthy economy makes people be able to afford more food (including onions).
  - A healthy economy boosts the stock market.

Please Don't...

…ban ice cream just to bring down the murder rate!

…make us eat tons of onions just to increase the stock market!

Photographs by Eink Newth and BrindleT.

Conclusion

- When doing correlations, make a scattergram first to get an idea of strength, direction, and outliers.
- Be careful when calculating r by hand.
- Beware of lurking variables – correlation is not necessarily causation!

Photo courtesy of Acf.

Chapter 4.2

Linear Regression and the Coefficient of Determination

Learning Objectives

At the end of this lecture, the student should be able to:

- Explain what the "least-squares line" is
- Identify and describe the components of the least-squares line equation
- Explain how to calculate the residuals
- Calculate and interpret the coefficient of determination (CD)

Introduction

- Least-squares line
- Least-squares line equation
- Dealing with prediction using the least-squares line
- Coefficient of Determination

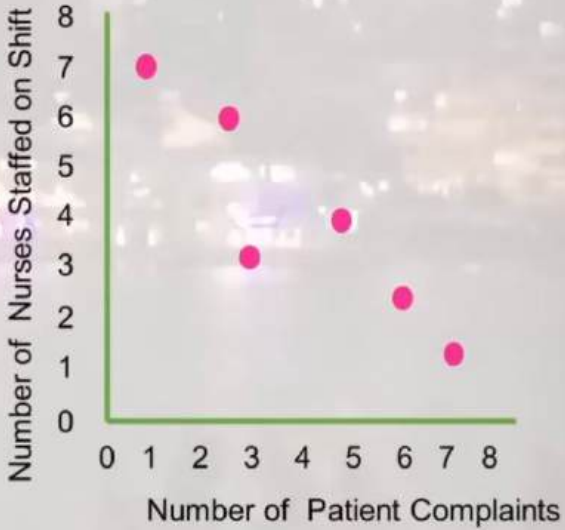
Painting in public domain

Least-Squares Criterion

What this means

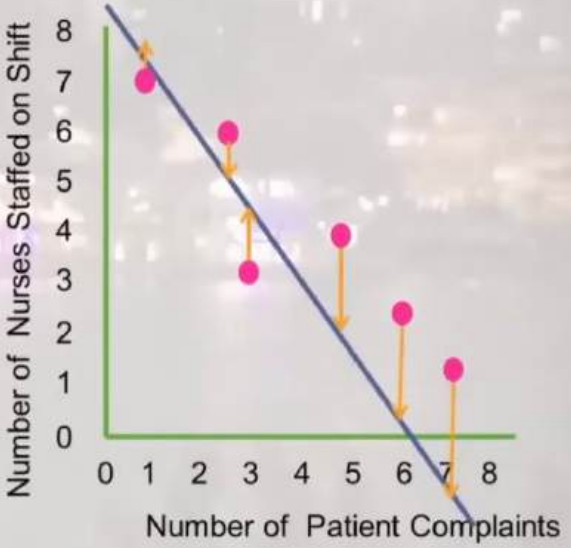
Where Does the Line Go?

- In the last chapter, we plotted scattergrams.
- I just drew a line for demonstration – but there is an official rule as to where this line goes.
- The rule is that the line has to meet the "least squares criterion"



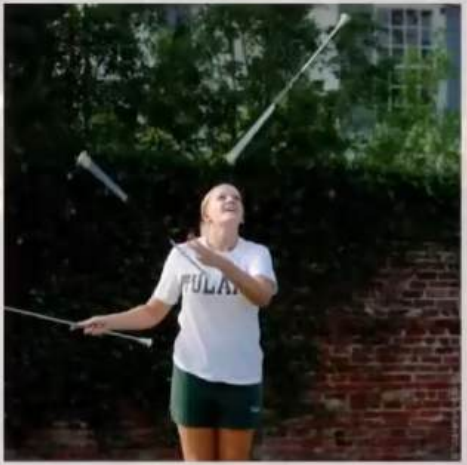
Where Does the Line Go?

- "least-squares line"
  - The vertical distances between the dot and line are squared to get rid of negative sign
  - These are called "squares"
- The line belongs where it would cause the smallest sum of squares for the whole dataset.



Where Does the Line Go?

- If you figure out where the line goes, you can draw it on a scatterplot. But how do you know exactly where it belongs on the graph?
- And what if you don't have a visual? How do you describe the line?
- You use an equation!



Picture courtesy of Tulane Public Relations

Least-Squares Line Equation

How to Find this Equation

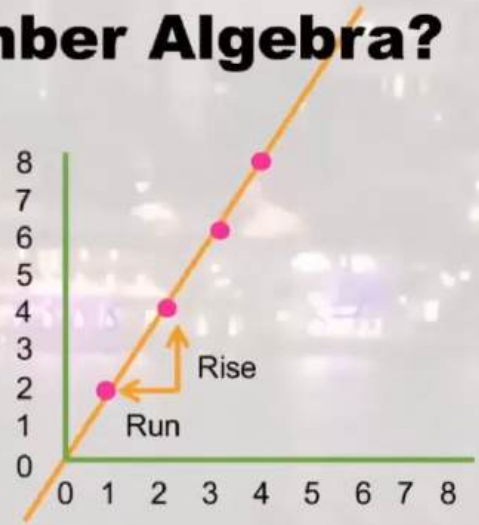
Remember Algebra?

x	y
1	2
2	4
3	6
4	8

Remember Algebra?

x	y
1	2
2	4
3	6
4	8

$y = bx + a$

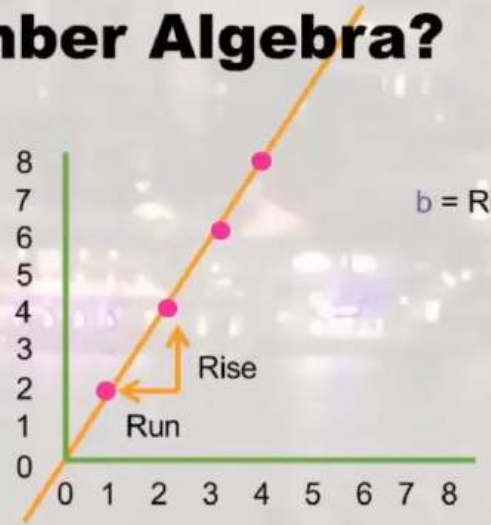


Remember Algebra?

x	y
1	2
2	4
3	6
4	8

$y = bx + a$

Slope



b = Rise/Run

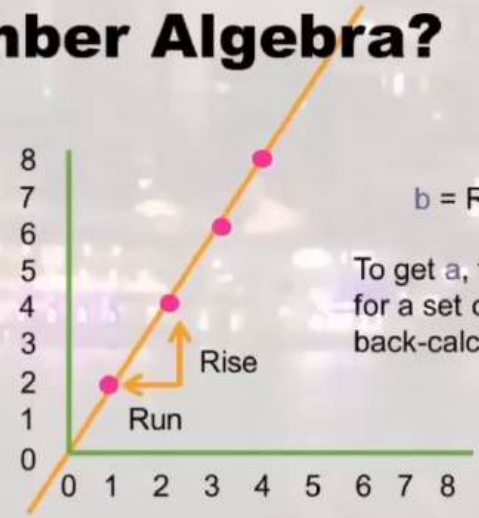
Remember Algebra?

x	y
1	2
2	4
3	6
4	8

$y = bx + a$

Slope

y-intercept

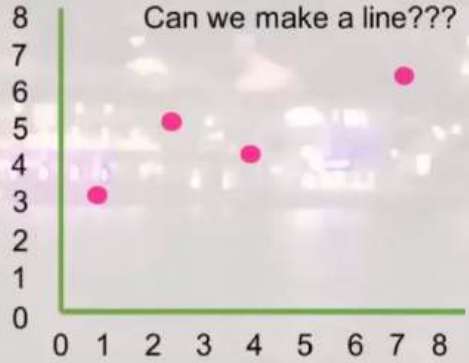


b = Rise/Run

To get a, fill in x and y for a set of x,y pairs and back-calculate.

Okay, now Statistics!

x	y
1	3
3	5
4	4
7	6



Okay, now Statistics!

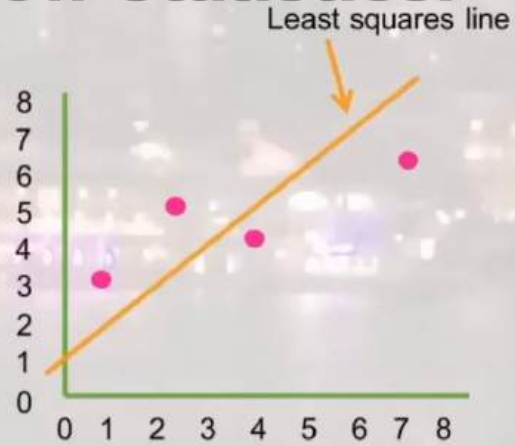
x	y
1	3
3	5
4	4
7	6

$$\hat{y} = bx + a$$

Hat (estimate)

Slope

y-intercept



Where Does the Line Go?

Software Approach

- Feed all the x,y pairs you have into the software.
- The software prints out the results in the form of an equation.
  - The "b" – slope
  - And the "a" – the y-intercept

Manual Approach (This Class!)

- Plug all the x,y pairs into an equation to get "b".
- Calculate x-bar and y-bar.
- Plug b, x-bar (for x), and y-bar (for y-hat) into the linear equation to back-calculate a.

Slope

$$\hat{y} = bx + a$$

Hat (estimate)

y-intercept

Recycling!

- Least-squares line is usually done along with r
- *SAVE YOUR CALCULATIONS* from r to recycle when calculating b:
  - Σx, Σy, Σx^2, and Σxy
- Also – save your r! You will need it later for the Coefficient of Determination.
- NOTE: You will need to calculate x-bar and y-bar – this was not done in r

Photograph by Patrick Nylin

x=DBP, y=# of Appointments

#	x	y	x^2	y^2	xy
1	70	3	4,900	9	210
2	115	45	13,225	2,025	5,175
3	105	21	11,025	441	2,205
4	82	7	6,724	49	574
5	93	16	8,649	256	1,488
6	125	62	15,625	3,844	7,750
7	88	12	7,744	144	1,056
	Σx = 678	Σy = 166	Σx^2 = 67,892	Σy^2 = 6,768	Σxy = 18,458

NOTE: Formula I'm using for b

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a = \overline{y} - \overline{bx}$$

GOAL: Fill in b and a so you have the least-squares line equation.

$$\hat{y} = bx + a$$

x=DBP, y=# of Appointments

#	x	y	x^2	y^2	xy
1	70	3	4,900	9	210
2	115	45	13,225	2,025	5,175
3	105	21	11,025	441	2,205
4	82	7	6,724	49	574
5	93	16	8,649	256	1,488
6	125	62	15,625	3,844	7,750
7	88	12	7,744	144	1,056
	Σx = 678	Σy = 166	Σx^2 = 67,892	Σy^2 = 6,768	Σxy = 18,458

NEW! x-bar = 678/7 = 96.9 y-bar = 166/7 = 23.7

NOTE: Formula I'm using for b

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a = \overline{y} - b\overline{x}$$

GOAL: Fill in b and a so you have the least-squares line equation.

$$\hat{y} = bx + a$$

x=DBP, y=# of Appointments

$n = 7$ $\quad \Sigma y = 166$

$\Sigma xy = 18{,}458$ $\quad \Sigma x^2 = 67{,}892$

$\Sigma x = 678$ $\quad \Sigma y^2 = 6{,}768$

$\overline{x} = 96.9$ $\quad \overline{y} = 23.7$

NOTE: Formula I'm using for b

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a = \overline{y} - b\overline{x}$$

GOAL: Fill in b and a so you have the least-squares line equation.

$$\hat{y} = bx + a$$

x=DBP, y=# of Appointments

$n = 7$ $\quad \Sigma y = 166$

$\Sigma xy = 18{,}458$ $\quad \Sigma x^2 = 67{,}892$

$\Sigma x = 678$ $\quad \Sigma y^2 = 6{,}768$

$\overline{x} = 96.9$ $\quad \overline{y} = 23.7$

$$b = \frac{(7)(18{,}458) - (678)(166)}{(7)(67{,}892) - (678)^2}$$

$$b = \frac{16{,}658}{15.560} = 1.1$$

NOTE: Formula I'm using for b

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a = \overline{y} - b\overline{x}$$

GOAL: Fill in b and a so you have the least-squares line equation.

$$\hat{y} = bx + a$$

x=DBP, y=# of Appointments

$n = 7$ $\quad \Sigma y = 166$

$\Sigma xy = 18{,}458$ $\quad \Sigma x^2 = 67{,}892$

$\Sigma x = 678$ $\quad \Sigma y^2 = 6{,}768$

$\overline{x} = 96.9$ $\quad \overline{y} = 23.7$

$b = 1.1$

$a = \overline{y} - b\overline{x}$

$a = 23.7 - (1.1 * 96.9)$

$a = -80.0$

NOTE: Formula I'm using for b

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a = \overline{y} - b\overline{x}$$

GOAL: Fill in b and a so you have the least-squares line equation.

$$\hat{y} = bx + a$$

x=DBP, y=# of Appointments

$n = 7$ $\quad \Sigma y = 166$

$\Sigma xy = 18{,}458$ $\quad \Sigma x^2 = 67{,}892$

$\Sigma x = 678$ $\quad \Sigma y^2 = 6{,}768$

$\overline{x} = 96.9$ $\quad \overline{y} = 23.7$

$b = 1.1$

$a = \overline{y} - b\overline{x}$

$a = 23.7 - (1.1 * 96.9)$

$a = -80.0$

CHECK! (1.1*96.9) – 80.0 should = 23.7!

NOTE: Formula I'm using for b

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a = \overline{y} - b\overline{x}$$

GOAL: Fill in b and a so you have the least-squares line equation.

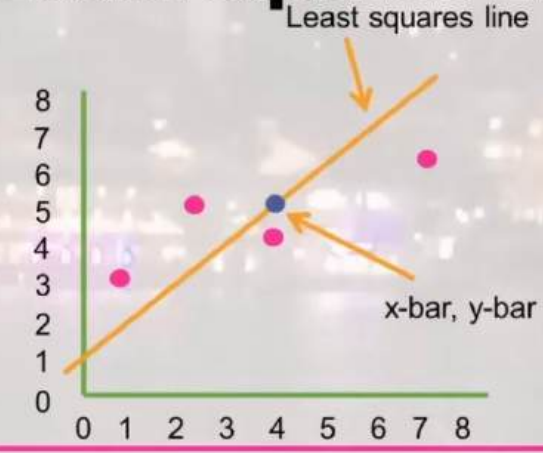
$$\hat{y} = 1.1x - 80.0$$

Predicting with the Least Squares Line Equation

Different Ways to Use the Equation

Rule About Least Squares Line

x	y
1	3
3	5
4	4
7	6



$\hat{y} = bx + a$

Hat (estimate)

Slope

y-intercept

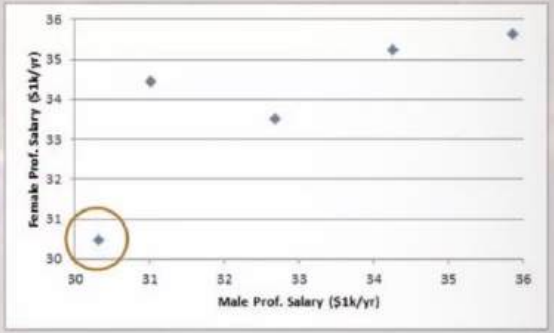
x-bar and y-bar *always* fall on the least squares line – but other points may or may not

Facts About the Slope (b)

- The slope (b) of the least-squares line tells us how many units the response variable (y) is expected to change for each 1 unit of change in the explanatory variable (x).
- For our example: $\hat{y} = 1.1x - 80.0$
  - x=DBP, y=# of Appointments
  - For each increase in 1 mmHg of DBP (x), there is a 1.1 increase in the number of appointments the patient had over the past year (y)
- The number of units change in the y for each unit change in x is called the "marginal change" in the y.

Influential Points

- Like with r, if a point is an outlier, it can drastically influence the least squares line equation.
- An extremely high x or extremely low x can do that.
- Always check the scattergram first for outliers!



What is the "Residual"?

- Once the equation is there, you can plug each x in, and get a y-hat out.
- Patient #1:
  - (1.1*70) – 80.0 = -3
- Patient #2:
  - (1.1*115) – 80.0 = 46.5

$\hat{y} = 1.1x - 80.0$

#	x	y
1	70	3
2	115	45

Residual is y minus y-hat

Patient #1: 3 – (-3) = 6

Patient #2: 45 – 46.5 = -1.5

Bottom Line: You don't want big residuals, because that would mean the line didn't fit very well.

Using Least Squares Line Equation for Prediction

- Let's say you knew someone's DBP and you wanted to predict how many appointments s/he would have next year
- You can plug the DBP in as x, and get y-hat out, and say that's your prediction
- If you use an x within the range of the original equation (70-125), this type of prediction is called interpolation.
- If you use an x from outside the range (such as 65, or 130), it is extrapolation – not a great idea.

#	x	y
1	70	3
2	115	45
3	105	21
4	82	7
5	93	16
6	125	62
7	88	12
	$\Sigma x =$ 678	$\Sigma y =$ 166

Example of Interpolation

The patient in your study has a DBP of 80. That is within the range of your x's. Let's predict how many appointments he will have next year. Here's the equation:

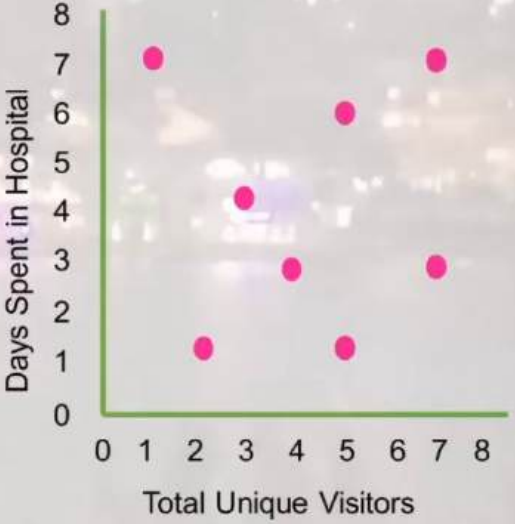
$\hat{y} = 1.1x - 80.0$

(1.1 * 80) – 80 = 8, so we predict this patient will come to 8 appointments next year.

#	x	y
1	70	3
2	115	45
3	105	21
4	82	7
5	93	16
6	125	62
7	88	12
	$\Sigma x =$ 678	$\Sigma y =$ 166

Is it Really This Easy to Make Predictions Using the Least Squares Line?

- No. You can make a linear equation out of any x,y pairs.
- If there is no linear correlation, though, the line is meaningless for prediction.
- Imagine a line for this scatter plot – would that really work for prediction?
- To evaluate if our least-squares line equation is should be used for interpretation, we use the Coefficient of Determination



Coefficient of Determination

Get out the r!

The Coefficient of Determination (CD)

- This is r^2 (in other words, r times r)
  - Then, like CV, we turn it into a %
- In the example, our r=0.95
- 0.95 * 0.95 = .90
- CD = 90%
- 90% = explained variation in y (by the linear equation)
- 100% - 90% = 10% unexplained variation

- "90% of the variation in the number of appointments is explained by DBP."
- "10% of the variation in the number of appointments is NOT explained by DBP."
- What happens if the CD is low?
  - CD should be better than at least 50% (random)
  - The higher, the better
  - If it is low, it means other variables might be needed to explain more of the variation

Chapter 4 Summary

- We started with quantitative x,y pairs
- We made a scatterplot to look at the linear relationship between x and y, and look at outliers
- We calculated r to see if our correlation was positive or negative, and weak, moderate or strong
- We calculated b and a to come up with the least-squares line equation
  - Notice: the sign on b will always match the sign on r (negative or positive)
  - Also notice: Strong correlations will give you high CDs
- We used the linear equation to calculate residuals
- We used r to calculate the CD to decide if we wanted to use the linear equation for prediction
- We decided it was good for prediction at 90%

Conclusion

- Least-squares criterion and calculating the least-squares line
- Reviewing issues with prediction using the least-squares line
- Coefficient of Determination (CD)

Photo courtesy of Fluzwup

Chapter 7.1

Normal Distribution & Empirical Rule

Learning Objectives

At the end of this lecture, the student should be able to:

- State two properties of the normal curve.
- State two differences between Chebyshev Intervals and the Empirical Rule
- Explain how to apply the Empirical Rule to a normal distribution

Introduction

- Remembering distributions
- Properties of the normal curve
- Remembering Chebyshev Intervals
- Empirical Rule
- Applying the Empirical Rule

Photograph by Mightyhansa

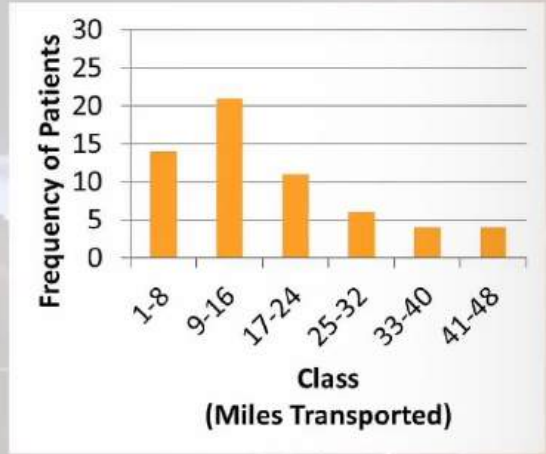
Remember Distributions?

- Using quantitative variable
- Classes determined
- Frequency table made

Class Limits	Freq-uency	Relative Freq-uency
1-8 miles	14	0.23
9-16 miles	21	0.35
17-24 miles	11	0.18
25-32 miles	6	0.10
33-40 miles	4	0.07
41-48 miles	4	0.07
Total	60	1.00

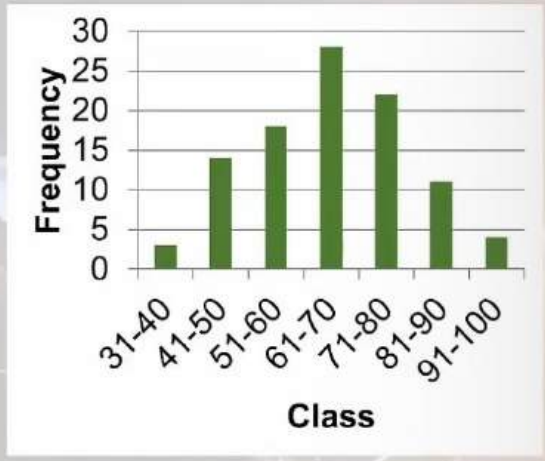
Remember Distributions?

- Using quantitative variable
- Classes determined
- Frequency table made
- Frequency histogram
- Then we could see the distribution



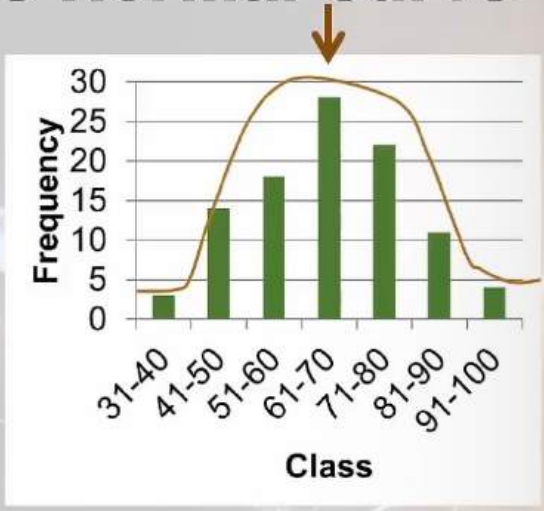
Remember the Normal Distribution?

- Imagine a large class (n=100) takes a very difficult test
- The test is worth 100 points, but it's so hard, no one actually gets 100 points
- Instead, the mode is near a C grade



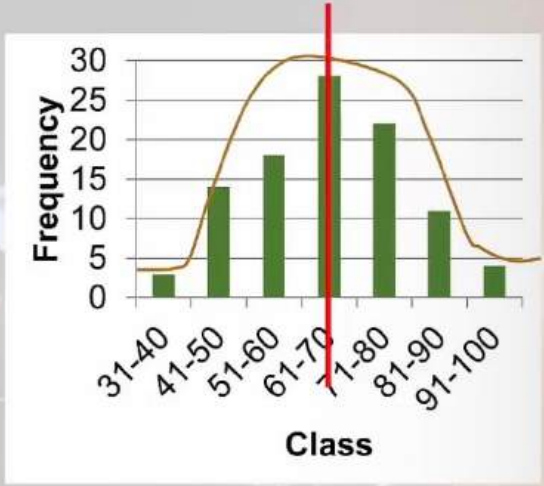
Properties of the Normal Curve

1. The curve is bell-shaped, with the highest point over the mean.



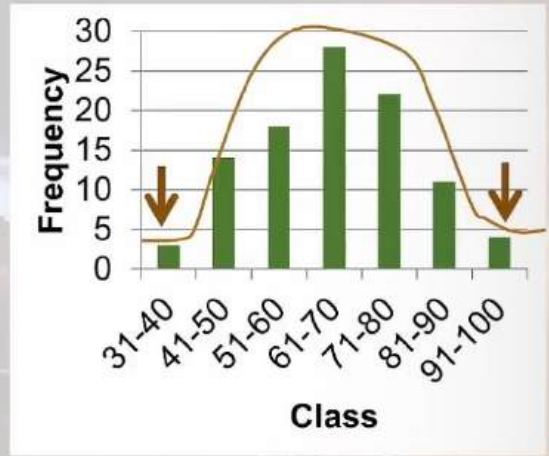
Properties of the Normal Curve

1. The curve is bell-shaped, with the highest point over the mean.
2. The curve is symmetrical around a vertical line through the mean.



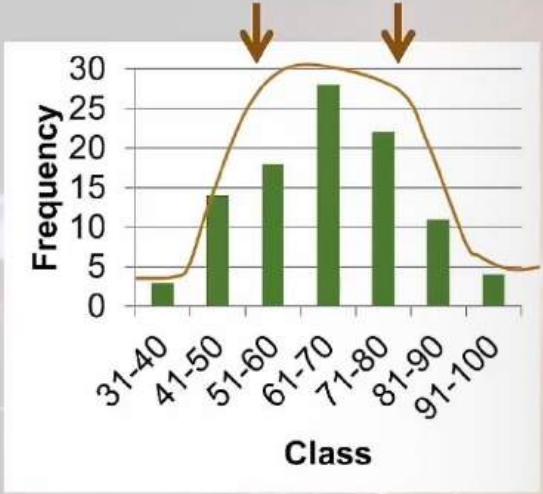
Properties of the Normal Curve

1. The curve is bell-shaped, with the highest point over the mean.
2. The curve is symmetrical around a vertical line through the mean.
3. The curve approaches the horizontal axis but never touches or crosses it.



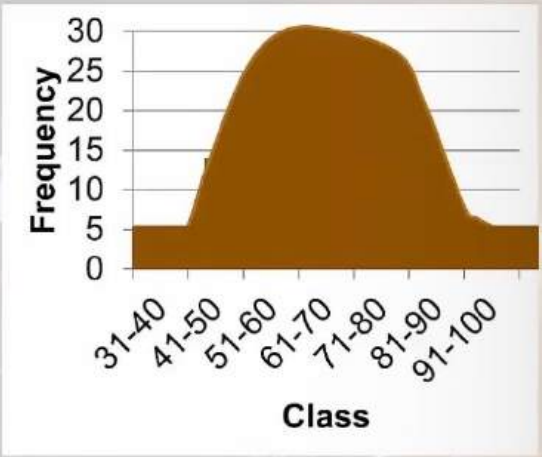
Properties of the Normal Curve

1. The curve is bell-shaped, with the highest point over the mean.
2. The curve is symmetrical around a vertical line through the mean.
3. The curve approaches the horizontal axis but never touches or crosses it.
4. The inflection (transition) points between cupping upward and downward occur at about mean +/- 1 sd



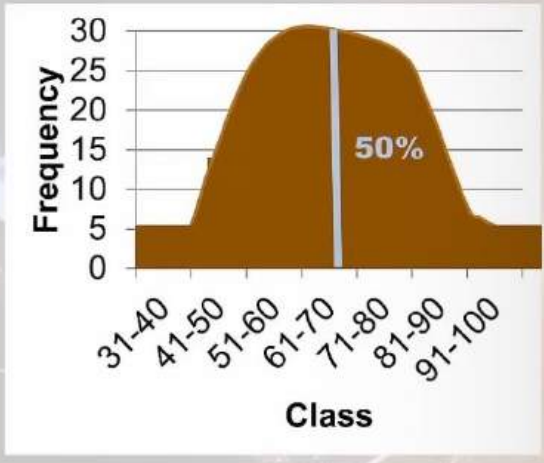
Properties of the Normal Curve

1. The curve is bell-shaped, with the highest point over the mean.
2. The curve is symmetrical around a vertical line through the mean.
3. The curve approaches the horizontal axis but never touches or crosses it.
4. The inflection (transition) points between cupping upward and downward occur at about mean +/- 1 sd
5. The area under the entire curve is 1 (think: 100%).



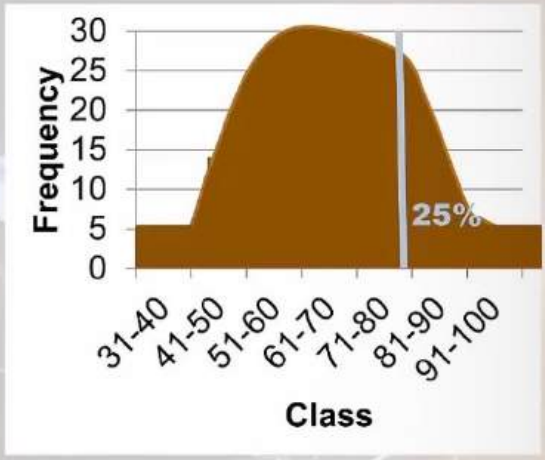
Properties of the Normal Curve

1. The curve is bell-shaped, with the highest point over the mean.
2. The curve is symmetrical around a vertical line through the mean.
3. The curve approaches the horizontal axis but never touches or crosses it.
4. The inflection (transition) points between cupping upward and downward occur at about mean +/- 1 sd
5. The area under the entire curve is 1 (think: 100%).



Properties of the Normal Curve

1. The curve is bell-shaped, with the highest point over the mean.
2. The curve is symmetrical around a vertical line through the mean.
3. The curve approaches the horizontal axis but never touches or crosses it.
4. The inflection (transition) points between cupping upward and downward occur at about mean +/- 1 sd
5. The area under the entire curve is 1 (think: 100%).



Empirical Rule

Remember Chebyshev?

Remember Chebyshev?

- Intervals have boundaries, or limits: lower limit and upper limit.
- Remember Chebyshev Intervals?
  - They say, "At least ____% of the data fall in the interval."
  - When lower limit was $\mu-2\sigma$, and upper limit was $\mu+2\sigma$, at least 75% of the data were in the interval.
- Imagine n=100 students, μ score on test 65.5, $\sigma = 14.5$
  - Lower limit: $65.5 - (2*14.5) = 36.5$
  - Upper limit: $65.5 + (2*14.5) = 94.5$
  - So if you had 100 data points, *at least* 75 would be between 36.5 and 94.5.

Chebyshev vs. Empirical Rule

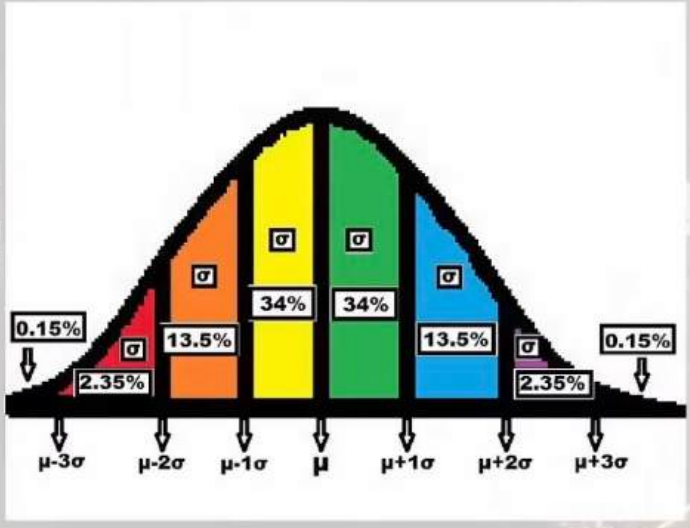
Chebyshev's Theorem

1. Applies to any distribution
2. Says "at least"
   - between μ +/- 2σ, there are *AT LEAST* 75% of the data
   - between μ +/- 3σ is at least 88.9%
   - between μ +/- 4σ is at least 93.8%

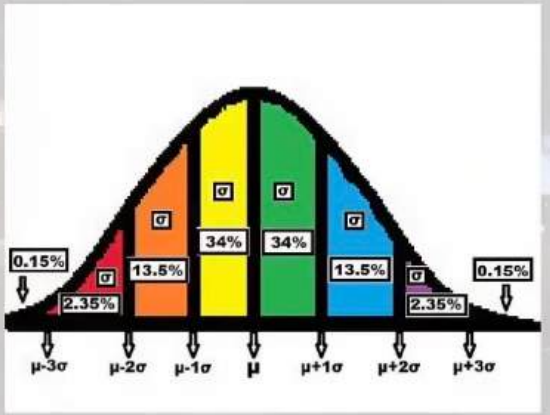
Empirical Rule

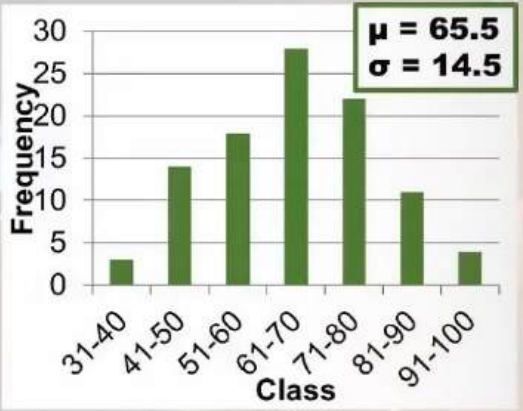
1. Applies to *ONLY* the normal distribution
2. Says "approximately"
   - 68% of the data are in interval μ +/- 1σ
   - 95% in interval μ +/- 2σ
   - 99.7% (almost all) in interval μ +/- 3σ

Empirical Rule Diagram

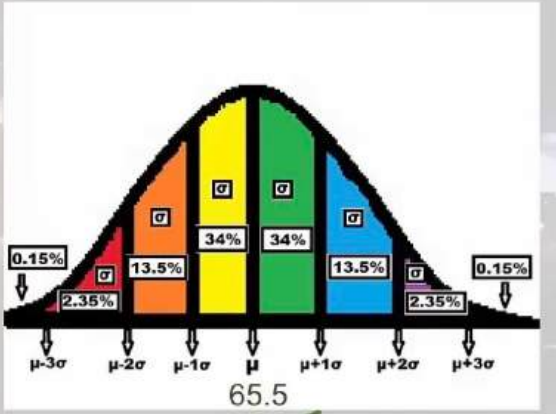


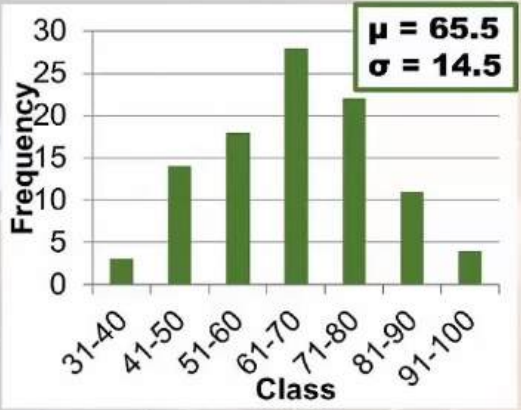
Applying Empirical Rule





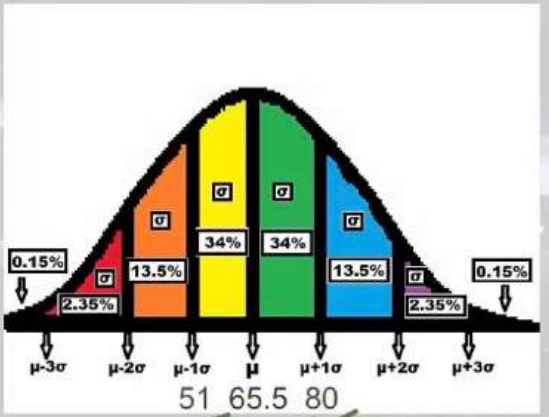
Applying Empirical Rule

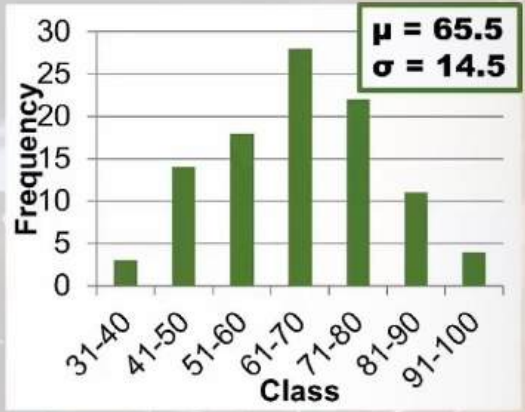




We can add the μ.

Applying Empirical Rule

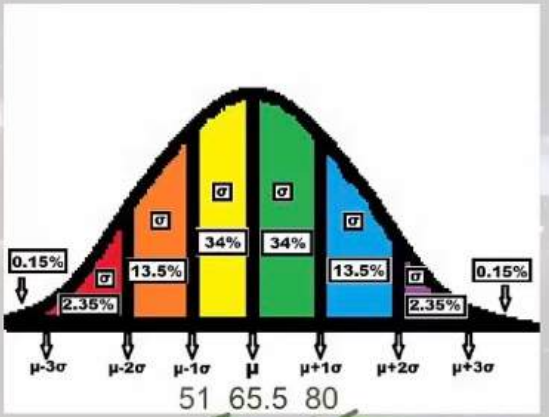


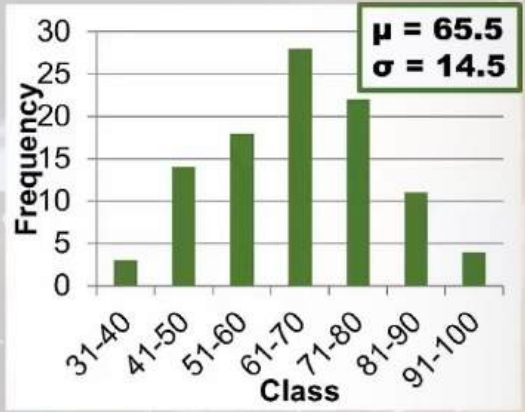


65.5 – (1*14.5) = 51

65.5 + (1*14.5) = 80

Applying Empirical Rule

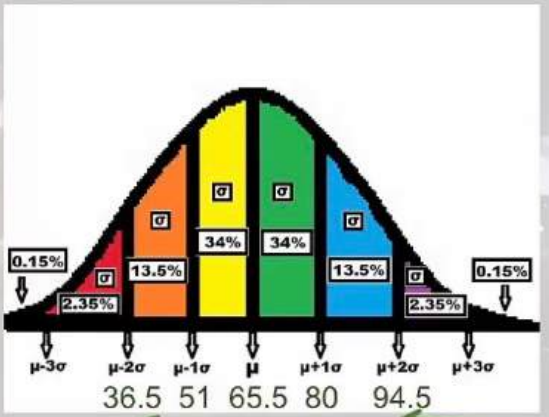


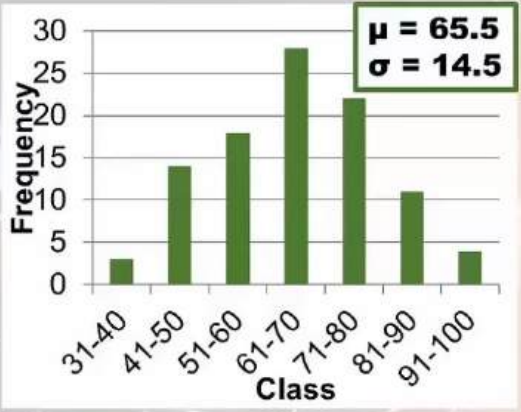


65.5 – (1*14.5) = 51

65.5 + (1*14.5) = 80

Applying Empirical Rule

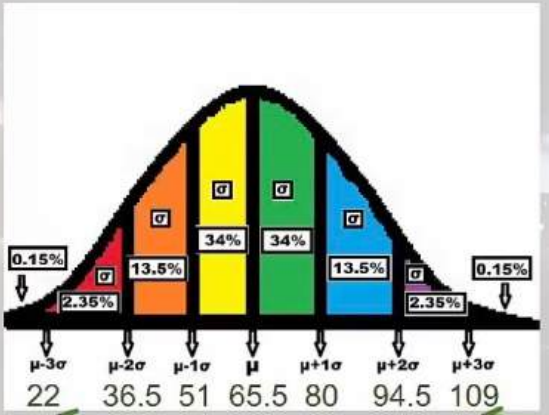


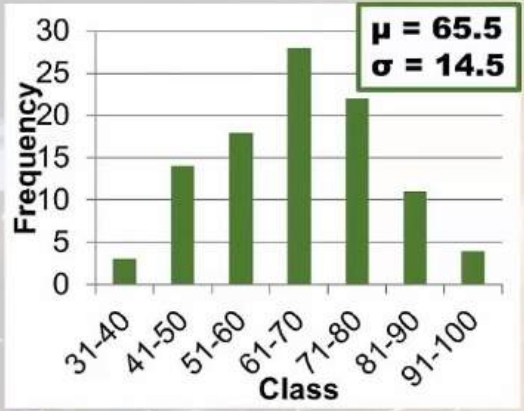


65.5 – (2*14.5) = 36.5

65.5 + (2*14.5) = 94.5

Applying Empirical Rule

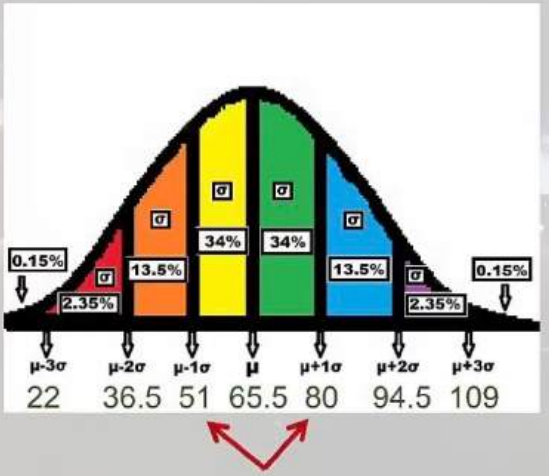




65.5 – (3*14.5) = 22

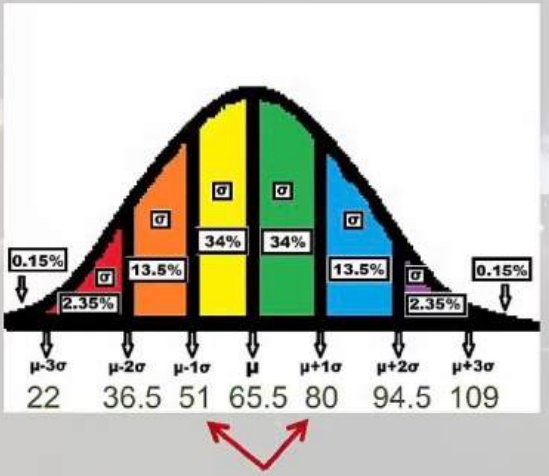
65.5 + (3*14.5) = 109

Applying Empirical Rule



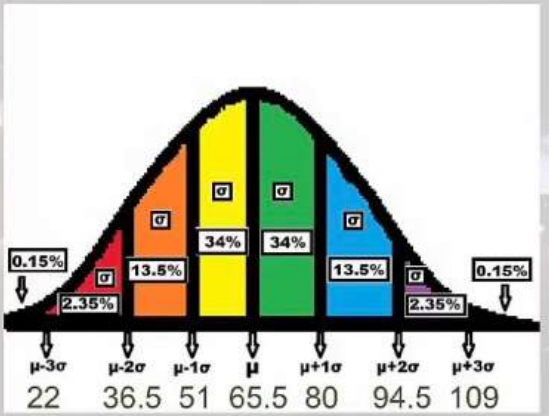
- Remember n=100?
- 34% of scores are between 51 and 65.5, meaning 34 scores.
- Another 34% (n=34) are between 65.5 and 80.
- This means 34%+34%=68% (n=68) are between 51 and 80.

Applying Empirical Rule



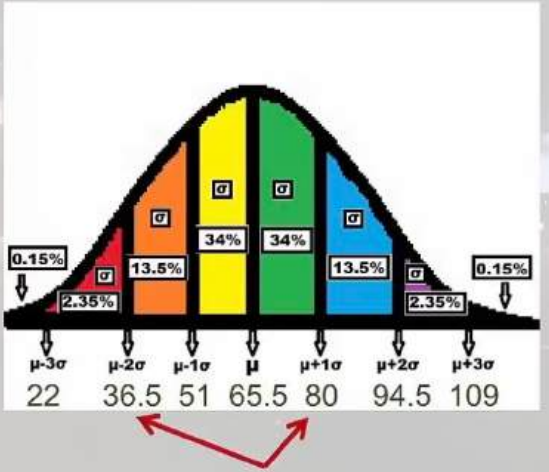
- Remember n=100?
- 34% of scores are between 51 and 65.5, meaning 34 scores.
- Another 34% (n=34) are between 65.5 and 80.
- This means 34%+34%=68% (n=68) are between 51 and 80.

Applying Empirical Rule



Question: What % of the data (student scores) are between 36.5 and 80?

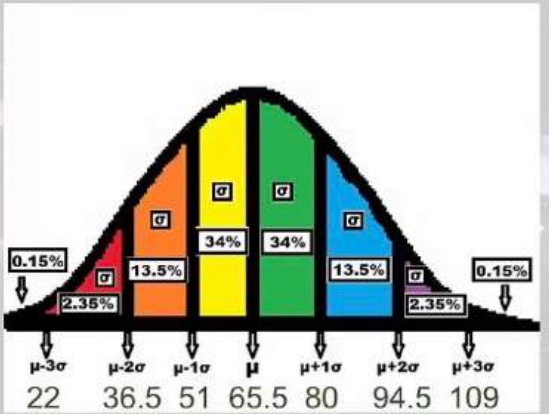
Applying Empirical Rule



Question: What % of the data (student scores) are between 36.5 and 80?

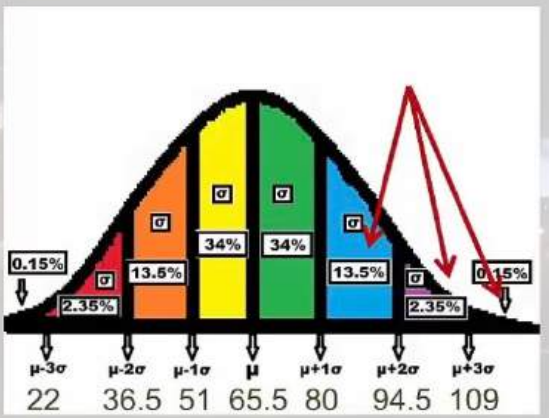
Answer: 13.5% + 34% + 34% = 81.5%

Applying Empirical Rule



Question: What cutpoint marks the top 16% of scores?

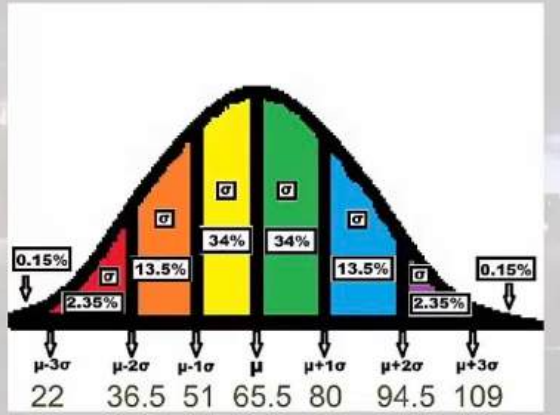
Applying Empirical Rule



Question: What cutpoint marks the top 16% of scores?

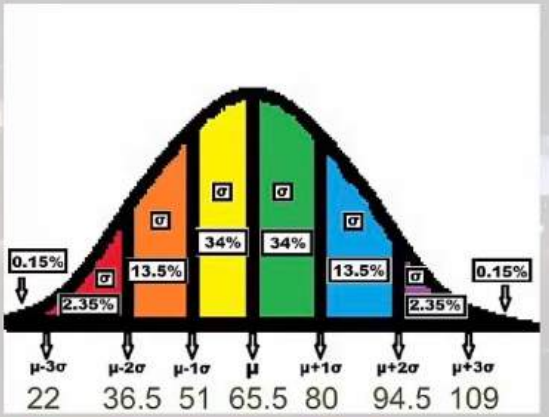
Answer: 0.15% + 2.35% + 13.5% = 16%. Top 16% cutpoint is 80.

Applying Empirical Rule



Question: What % of scores are below 94.5?

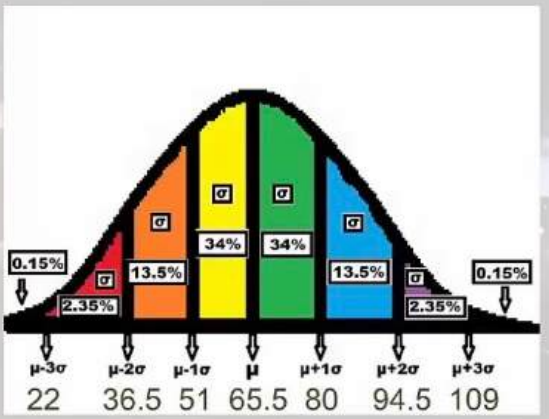
Applying Empirical Rule



Question: What % of scores are below 94.5?

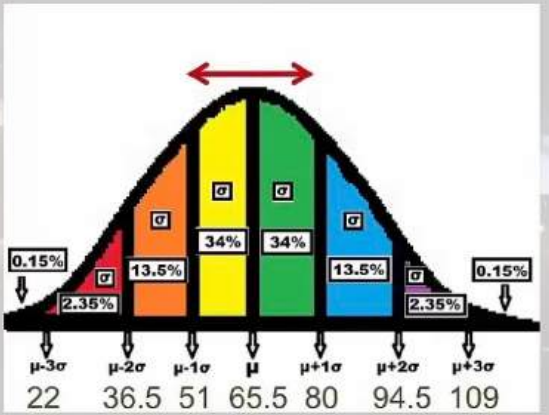
Answer: Add up % below 94.5: 13.5% + 34% + 34% + 13.5% + 2.35% + 0.15% = 97.5%

Applying Empirical Rule



Question: What are the cutpoints for the middle 68% of the data?

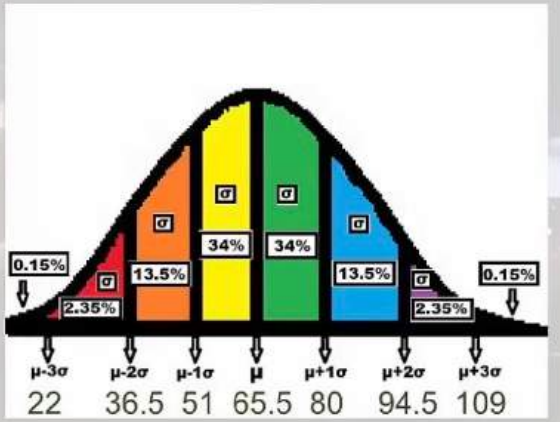
Applying Empirical Rule



Question: What are the cutpoints for the middle 68% of the data?

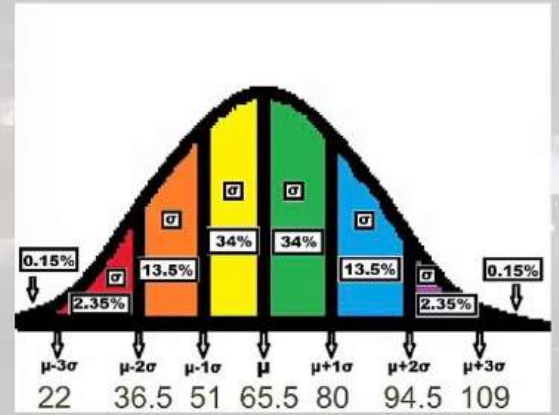
Answer: Middle 68% means 34% above mean (80) and 34% below mean (51).

Applying Empirical Rule



Question: What is the *probability* that if I select one student, that student will have a score less than 80?

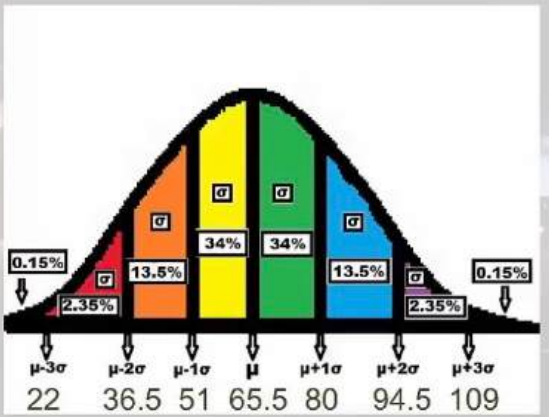
Applying Empirical Rule



Question: What is the *probability* that if I select one student, that student will have a score less than 80?

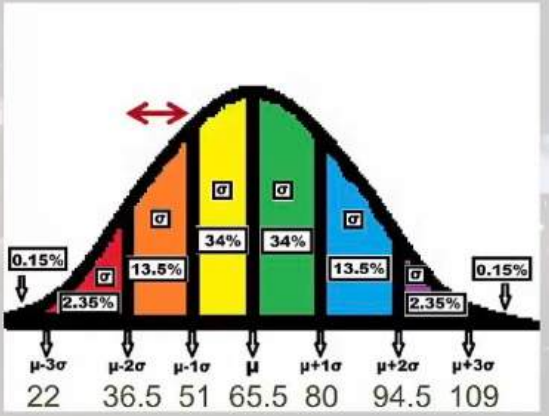
Answer: 50% + 34% = 84%

Applying Empirical Rule



Question: What is the *probability* I will select a student with a score between 36.5 and 51?

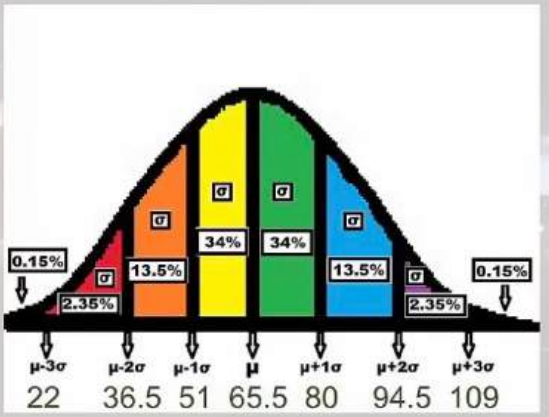
Applying Empirical Rule



Question: What is the *probability* I will select a student with a score between 36.5 and 51?

Answer: 13.5%

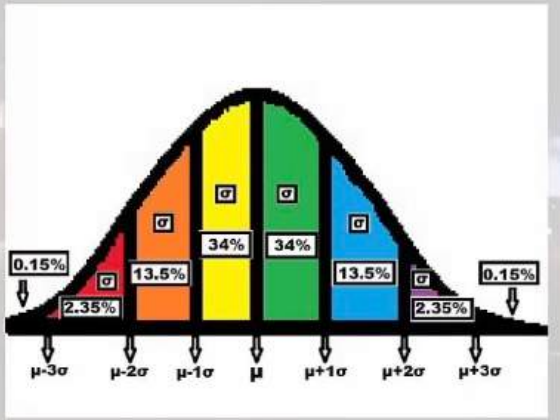
Applying Empirical Rule



Think about what would happen in a different class taking the same hard test.

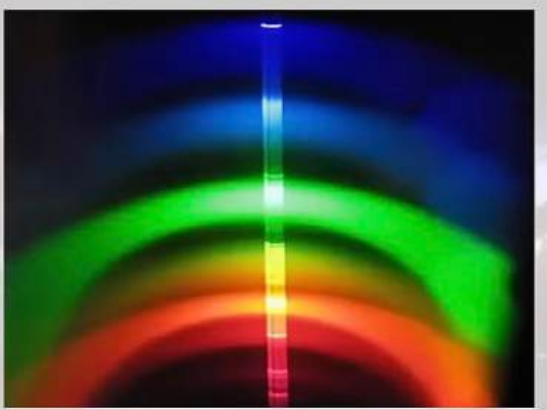
- What if the μ was the same, but the σ was larger than 14.5? What would that do to the intervals?
- What if σ was smaller than 14.5?

%, Area, and Probability: Did you Know?



- The %'s literally refer to the % of the area of the shape.
- Imagine the whole thing as 100%.
- Example:
  - The orange part is 13.5% of the area
  - It is also the probability that an "x" between μ-1σ and μ-2σ will be selected

Conclusion



- The Empirical Rule helps establish intervals that apply to normally distributed data
- These intervals have a certain percentage of the data points in them
- These intervals depend on the mean and standard deviation of the data distribution

Image by Zaereth

Chapter 7.2 & 7.3

Z-scores & Probabilities

Learning Objectives

At the end of this lecture, the student should be able to:

- Explain how to convert an x to a z-score.
- Show how to look up a z-score in a z table.
- Explain how to find the probability of an x falling between two values on a normal distribution.
- Describe how to use the z table to look up a z corresponding to a percentage.
- Describe how to use the formula to calculate x from a z-score

Introduction

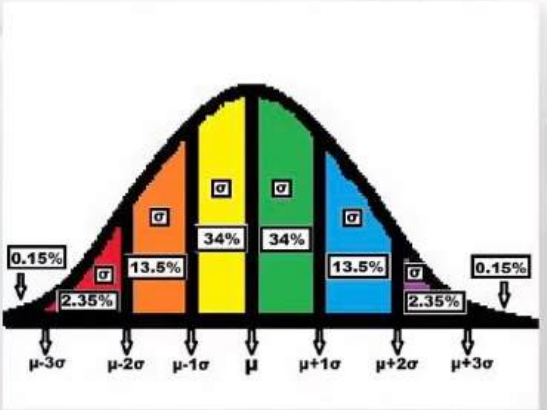
- Z-score and the standard normal distribution
- Z-score probabilities
- Using the z table to answer harder questions
- Calculating x from z
- Using z-score and probabilities correctly

Photograph by Dirk Beyer

What is a Z-Score?

Introduction to Standard Normal Distribution

Remember the Empirical Rule?

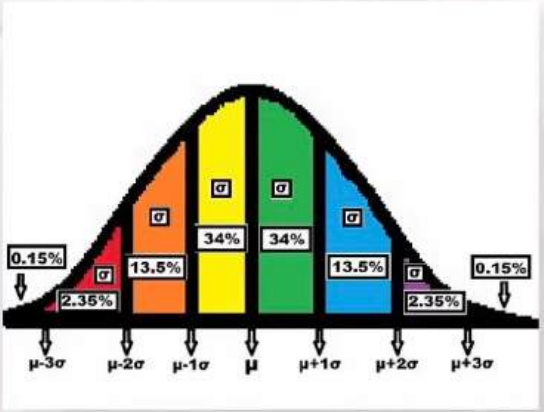


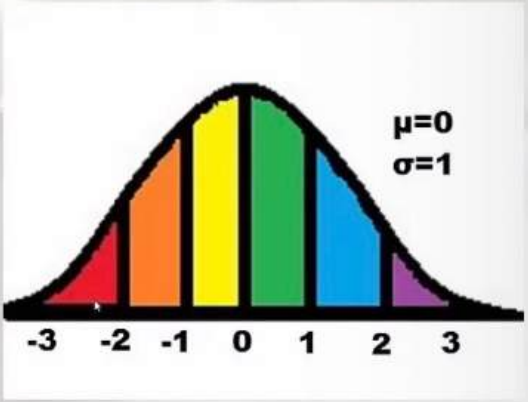
- Required normal distribution
- Worked well for the cutpoints available
  - What about in-between?

Remember the Empirical Rule?

- Required normal distribution
- Worked well for the cutpoints available
  - What about in-between?
- Notice the numbers along the bottom
  - -3, -2, -1, then μ (which has no number, or 0), then 1, 2, 3

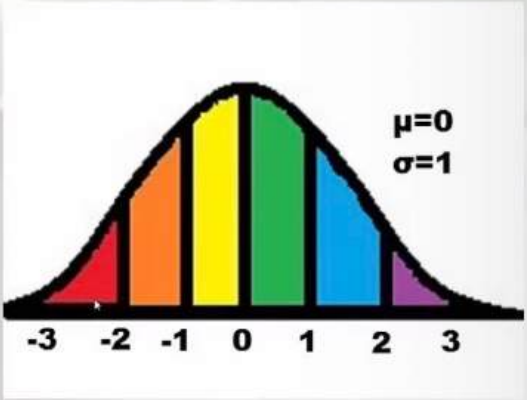
Z is the Standard Normal Distribution





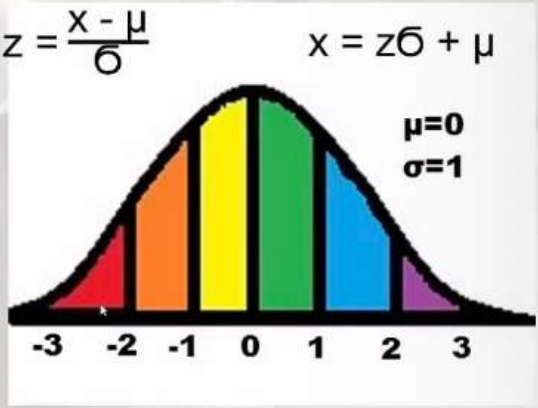
Z-scores

- Every value on a normal distribution (every "x") can be converted to a z-score.
- You must know the following to use formula:
  - The "x" – what you want to convert to z
  - The μ of the distribution
  - The σ of the distribution



Z-scores

- Every value on a normal distribution (every "x") can be converted to a z-score.
- You must know the following to use formula:
  - The "x" – what you want to convert to z
  - The μ of the distribution
  - The σ of the distribution

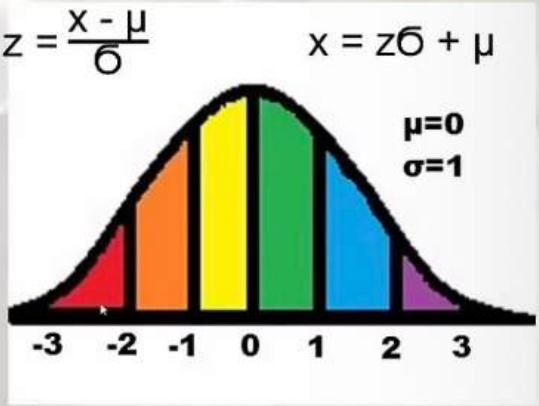


Z-scores: Smart Friend Example

$\mu = 65.5$
$\sigma = 14.5$

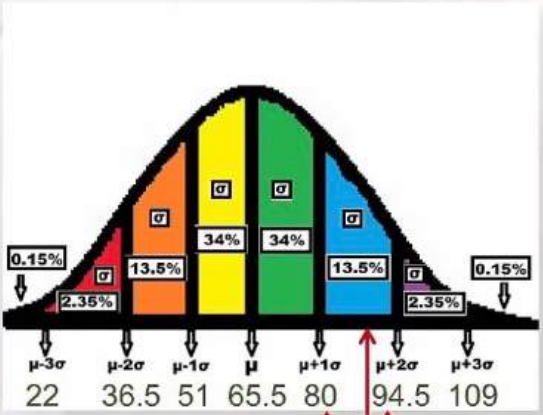
- Remember our n=100 students?
- Let's say your friend got a 90. What is the z-score for 90?

$$z = \frac{x - \mu}{\sigma} \qquad x = z\sigma + \mu$$

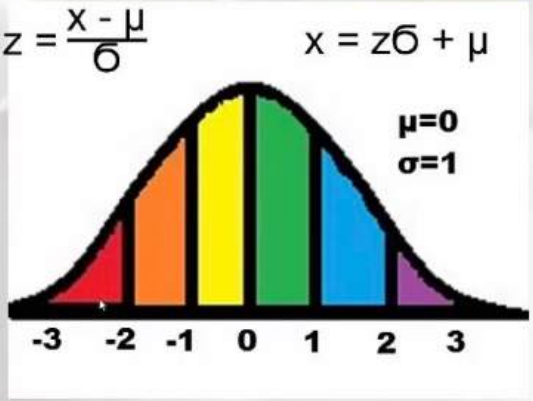


Z-scores: Smart Friend Example

$\mu = 65.5$
$\sigma = 14.5$



$$z = \frac{x - \mu}{\sigma} \qquad x = z\sigma + \mu$$



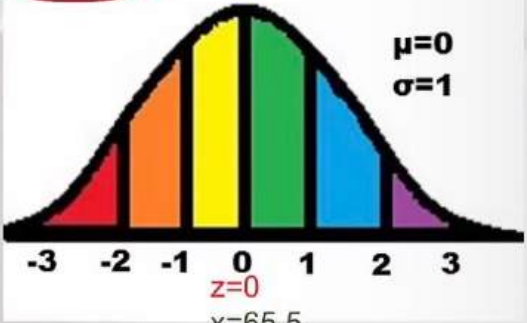
Z-scores: Smart Friend Example

$\mu = 65.5$
$\sigma = 14.5$

- Remember our n=100 students?
- Let's say your friend got a 90. What is the z-score for 90?
  - x=90
  - $\mu = 65.5$
  - $\sigma = 14.5$

$$z = \frac{x - \mu}{\sigma}$$

$$x = z\sigma + \mu$$



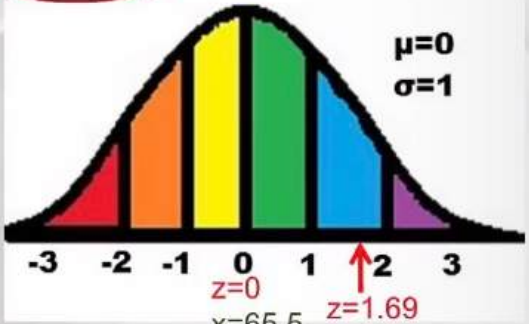
Z-scores: Smart Friend Example

$\mu = 65.5$
$\sigma = 14.5$

- Remember our n=100 students?
- Let's say your friend got a 90. What is the z-score for 90?
  - x=90
  - $\mu = 65.5$
  - $\sigma = 14.5$
- (90-65.5)/14.5 = 1.69

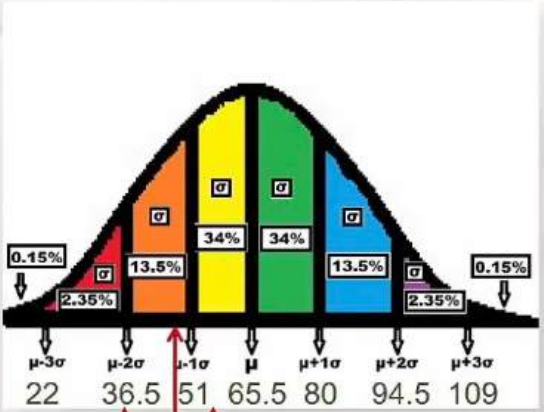
$$z = \frac{x - \mu}{\sigma}$$

$$x = z\sigma + \mu$$

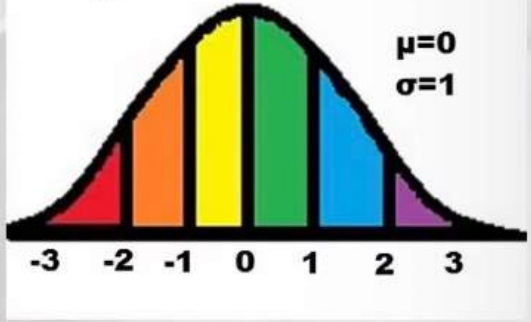


Z-scores: Not-so-smart Friend Example

$\mu = 65.5$
$\sigma = 14.5$



$$z = \frac{x - \mu}{\sigma} \qquad x = z\sigma + \mu$$



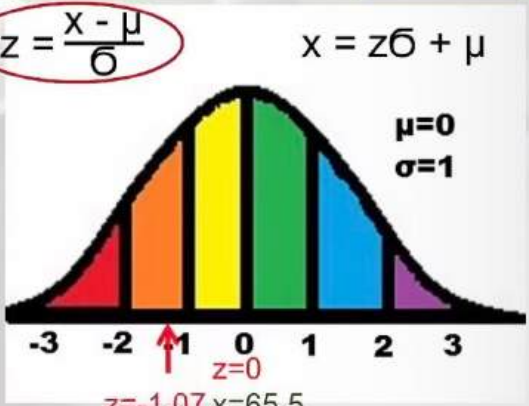
Z-scores: Not-so-smart Friend Example

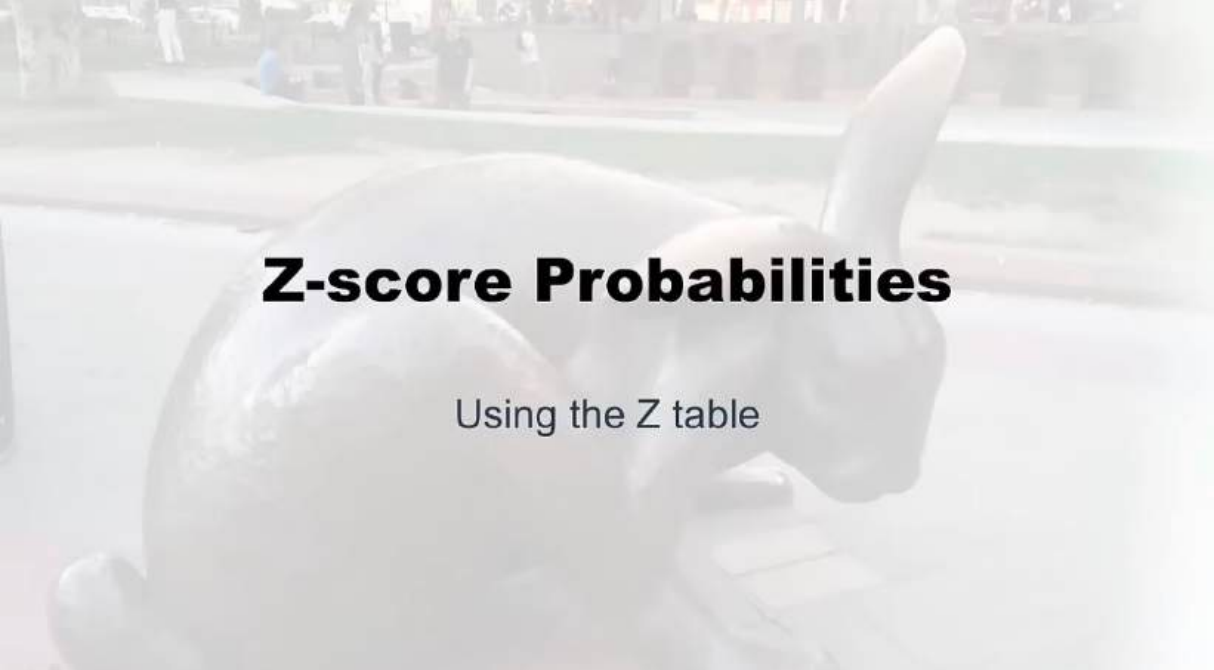
$\mu = 65.5$
$\sigma = 14.5$

- n=100
- Let's say your other friend got a 50. What is the z-score for 50?
  - x=50
  - $\mu = 65.5$
  - $\sigma = 14.5$
- (50-65.5)/14.5 = -1.07
- It is negative because it is below μ

$$z = \frac{x - \mu}{\sigma}$$

$$x = z\sigma + \mu$$

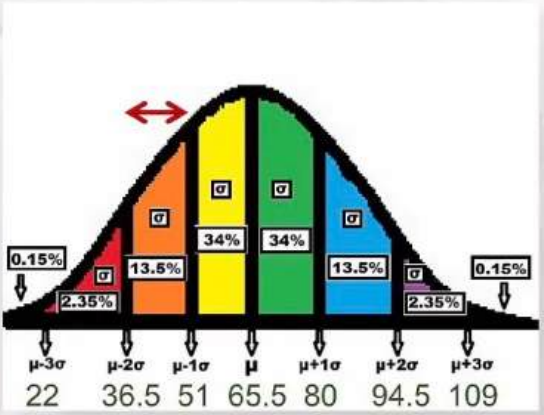




Z-score Probabilities

Using the Z table

Remember "Probability" from the Empirical Rule?



Question: What is the *probability* I will select a student with a score between 36.5 and 51?

Answer: 13.5%

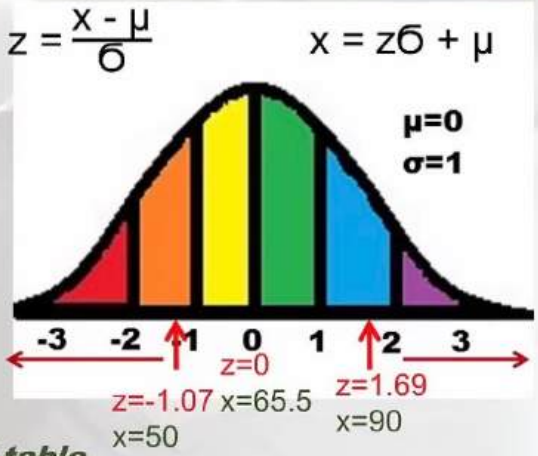
But what if you have z-scores of 1.69 (smart friend) and -1.07 (not-so-smart friend)?

Questions about Z-Score Probabilities

$\mu = 65.5$
$\sigma = 14.5$

- What is the *probability* that students scored *above* the smart friend?
  - In other words – what is the area under the curve from z=1.69 all the way up?
- What is the *probability* that students scored *below* the not-so-smart friend?
  - In other words – what is the area under the curve from z=-1.07 all the way down?

$$z = \frac{x - \mu}{\sigma} \qquad x = z\sigma + \mu$$

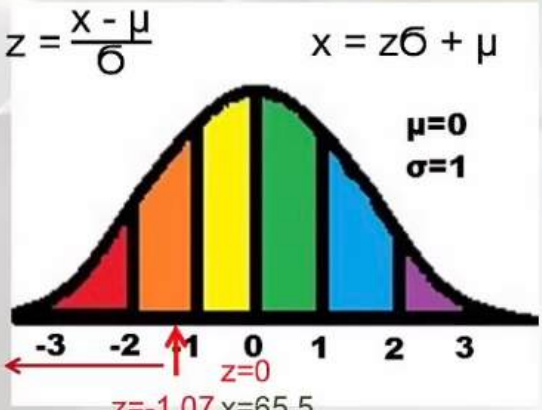


We will look these up using the Z table.

How to Use the Z Table

$\mu = 65.5$
$\sigma = 14.5$

- First, figure out what area you want.
  - What is the *probability* that students scored *below* the not-so-smart friend (z=-1.07, x=50)?
- *For areas to the left of a specified z value, use the table entry directly.*



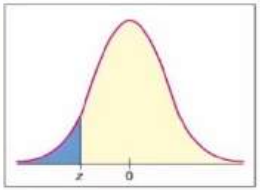


Table entry for z is the area to the left of z.

TABLE 4 Areas of a Standard Normal Distribution

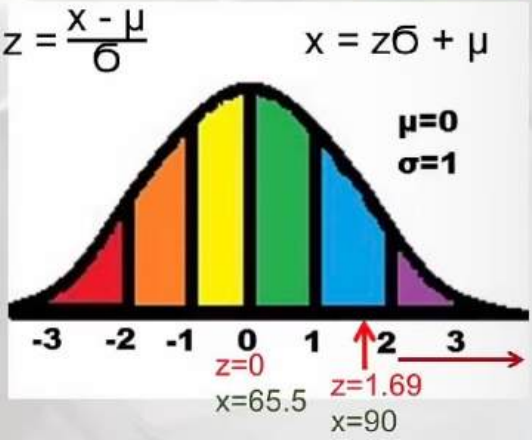
(a) Table of Areas to the Left of z

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183

How to Use the Z Table

$\mu = 65.5$
$\sigma = 14.5$

- Let's do the smart friend's probability.
  - What is the *probability* that students scored *above* the smart friend (x=90, z=1.69)?
- *For areas to the right of a specified z value, either:*
  - *Look up in table, then subtract result from 1, or*
  - *Use the opposite z (-1.69)*



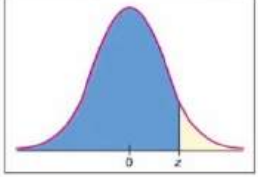


Table entry for z is the area to the left of z.

TABLE 4(a) *continued*

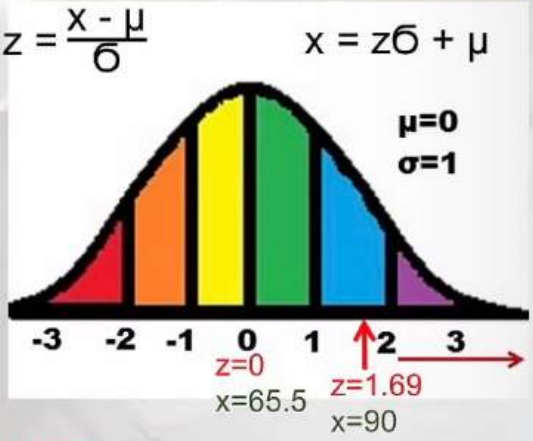
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857

How to Use the Z Table

$\mu = 65.5$
$\sigma = 14.5$

- Let's do the smart friend's probability.
  - What is the *probability* that students scored *above* the smart friend (x=90, z=1.69)?
- *For areas to the right of a specified z value, either:*
  - *Look up in table, then subtract result from 1, or*
  - *Use the opposite z (-1.69)*

$z = \frac{x - \mu}{\sigma}$ $x = z\sigma + \mu$

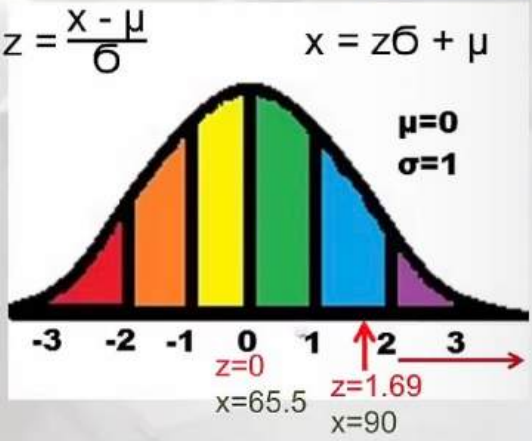


p = 0.9545. 1 – 0.9545 = 0.0455, or 4.55%

How to Use the Z Table

$\mu = 65.5$
$\sigma = 14.5$

- Let's do the smart friend's probability.
  - What is the *probability* that students scored *above* the smart friend (x=90, z=1.69)?
- *For areas to the right of a specified z value, either:*
  - *Look up in table, then subtract result from 1, or*
  - *Use the opposite z (-1.69)*



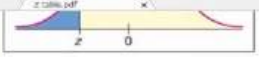


Table entry for z is the area to the left of z.

−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867

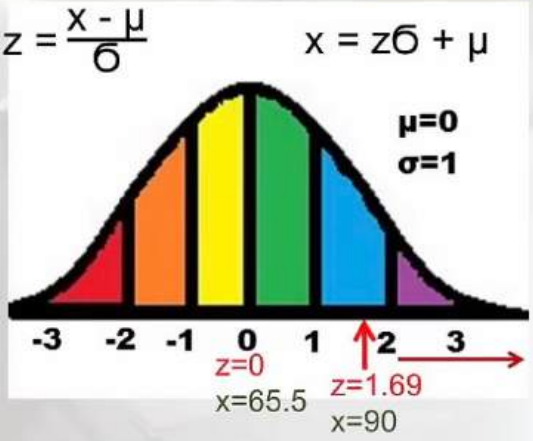
How to Use the Z Table

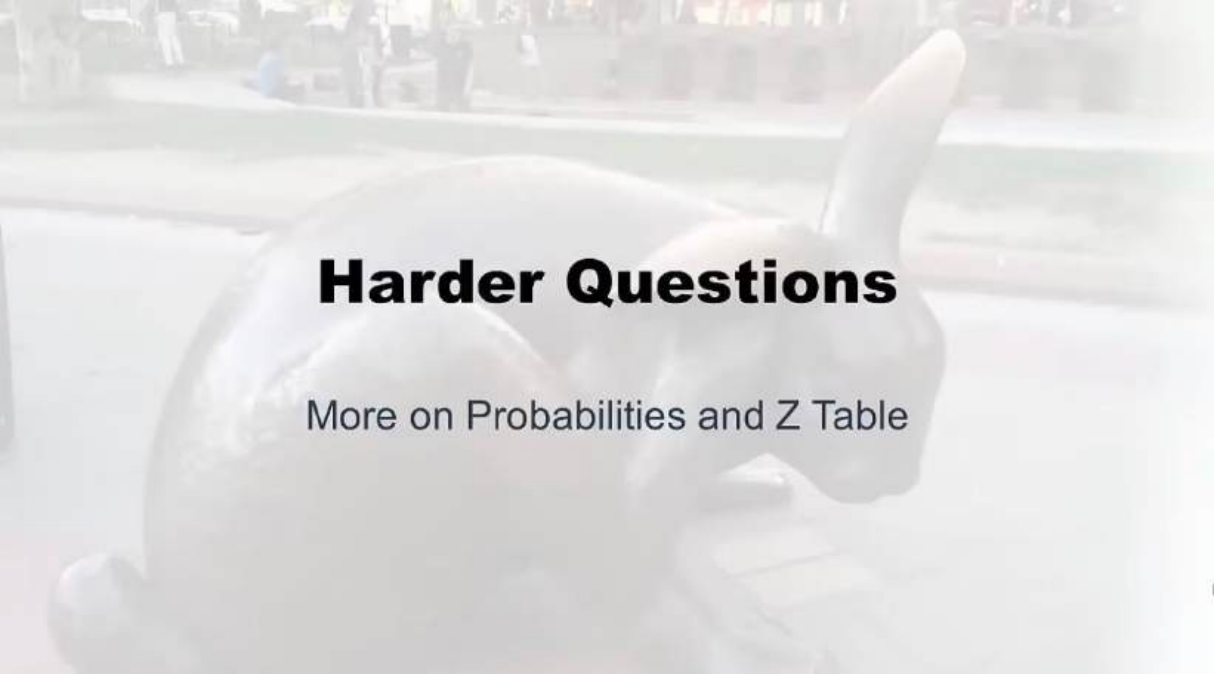
$\mu = 65.5$
$\sigma = 14.5$

- Let's do the smart friend's probability.
  - What is the *probability* that students scored *above* the smart friend (x=90, z=1.69)?
- *For areas to the right of a specified z value, either:*
  - *Look up in table, then subtract result from 1, or*
  - *Use the opposite z (-1.69)*

$p = 0.0455$, or 4.55%

$z = \frac{x - \mu}{\sigma}$ $\quad x = z\sigma + \mu$





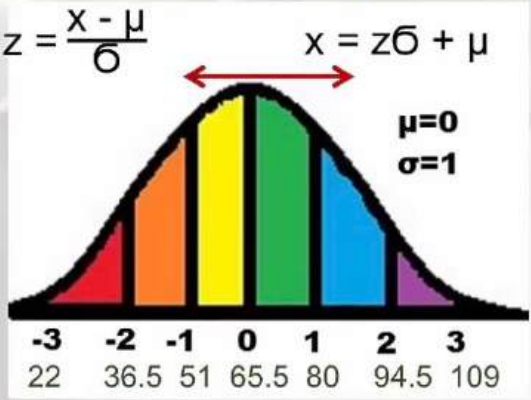
Harder Questions

More on Probabilities and Z Table

Harder Questions

$\mu = 65.5$
$\sigma = 14.5$

- What if you are looking at a *probability between two scores* – such as the probability the students will score between 50 and 90?

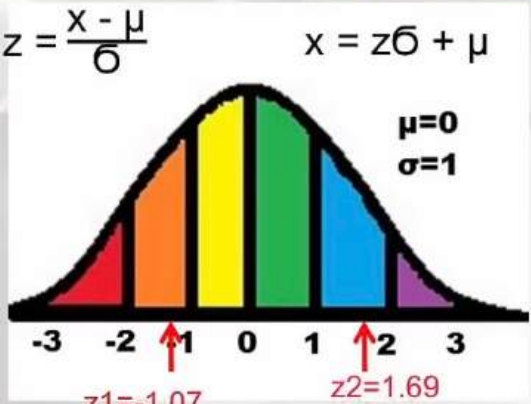


Calculating Probability Between Scores

$\mu = 65.5$
$\sigma = 14.5$

1. Note that you have x1 and x2 (two x's)
2. Calculate z1 and z2

$$z = \frac{x - \mu}{\sigma} \qquad x = z\sigma + \mu$$

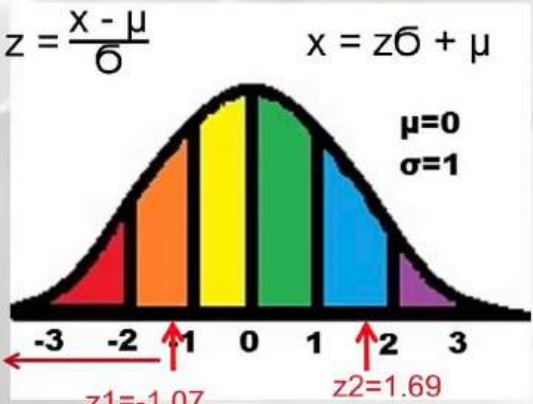


Calculating Probability Between Scores

$\mu = 65.5$
$\sigma = 14.5$

1. Note that you have x1 and x2 (two x's)
2. Calculate z1 and z2
3. For z1, find the probability to the *left (below)* z
   - Remember, direct probability from z table

$$z = \frac{x - \mu}{\sigma} \qquad x = z\sigma + \mu$$

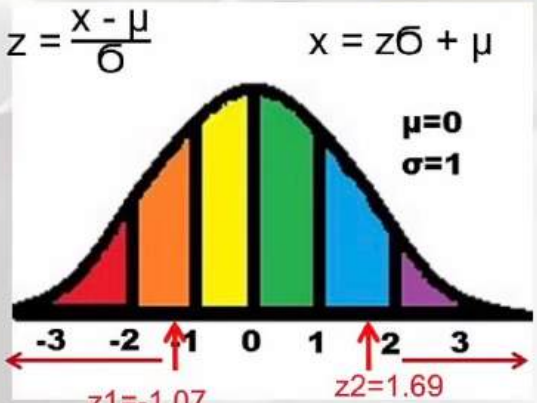


Calculating Probability Between Scores

$\mu = 65.5$
$\sigma = 14.5$

1. Note that you have x1 and x2 (two x's)
2. Calculate z1 and z2
3. For z1, find the probability to the *left (below)* z
   - Remember, direct probability from z table
4. For z2, find the probability to the *right (above)* z
   - Use one of the 2 methods

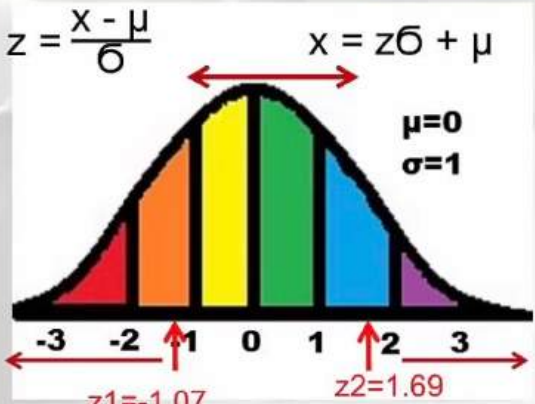
$$z = \frac{x - \mu}{\sigma} \qquad x = z\sigma + \mu$$



Calculating Probability Between Scores

μ = 65.5
σ = 14.5

1. Note that you have x1 and x2 (two x's)
2. Calculate z1 and z2
3. For z1, find the probability to the *left (below)* z
   - Remember, direct probability from z table
4. For z2, find the probability to the *right (above)* z
   - Use one of the 2 methods
5. Subtract both z1 and z2 probabilities from 1.

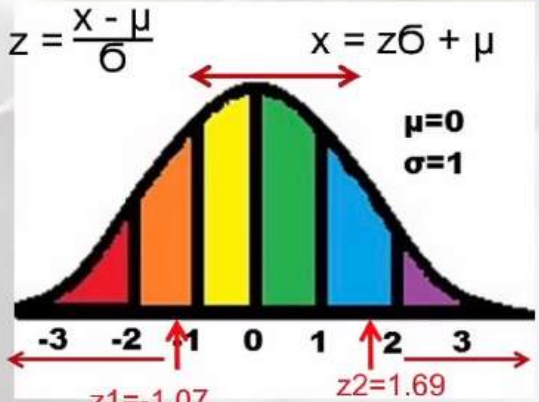


1 – z1 probability – z2 probability = between probability

Calculating Probability Between Scores

μ = 65.5
σ = 14.5

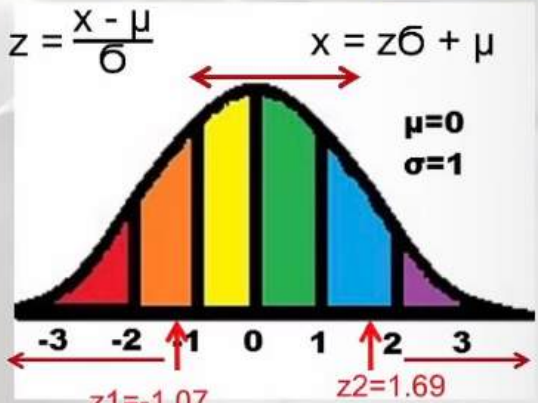
1. Note that you have x1 and x2 (two x's)
2. Calculate z1 and z2
3. For z1, find the probability to the *left (below)* z
   - It was *p = 0.1423.*



Calculating Probability Between Scores

$\mu = 65.5$
$\sigma = 14.5$

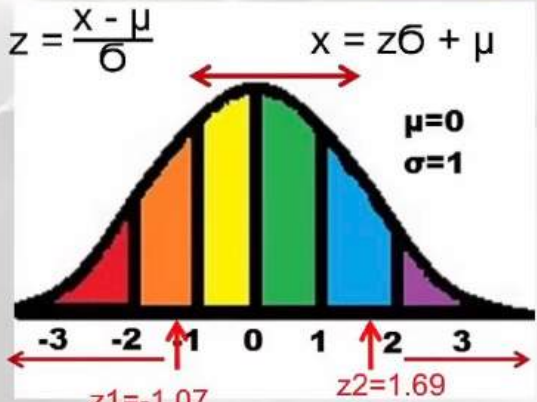
1. Note that you have x1 and x2 (two x's)
2. Calculate z1 and z2
3. For z1, find the probability to the *left (below)* z
   - It was $p = 0.1423$.
4. For z2, find the probability to the *right (above)* z
   - Use one of the 2 methods



Calculating Probability Between Scores

μ = 65.5
σ = 14.5

1. Note that you have x1 and x2 (two x's)
2. Calculate z1 and z2
3. For z1, find the probability to the *left (below)* z
   - It was *p = 0.1423.*
4. For z2, find the probability to the *right (above)* z
   - Use one of the 2 methods
5. Subtract both z1 and z2 probabilities from 1.



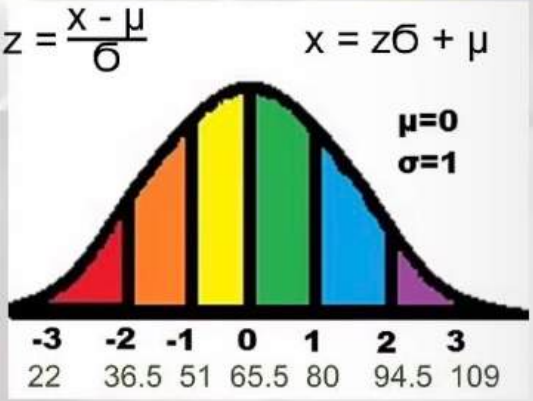
1 – 0.1423 – 0.0455 = 0.8122, or 81.22%

Harder Questions

$\mu = 65.5$
$\sigma = 14.5$

- What if you are looking at a probability more than 50%?
  - such as the probability students will score *greater than 50?*

$$z = \frac{x - \mu}{\sigma} \qquad x = z\sigma + \mu$$



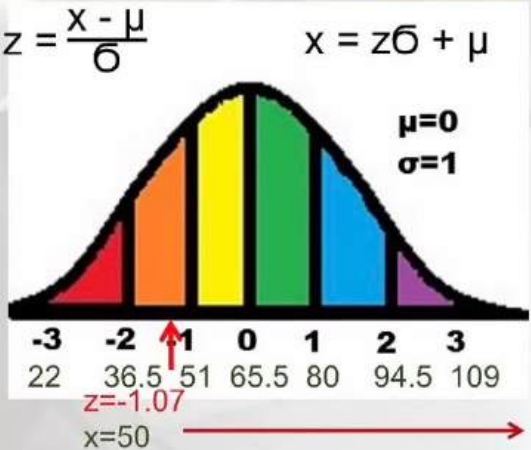
Harder Questions

$\mu = 65.5$
$\sigma = 14.5$

- What is the probability students will score *greater than 50?*
- *For areas to the right of a specified z value, either:*
  - *Look up in table, then subtract result from 1, or*
  - *Use the opposite z (1.07)*

Method 1: 1 - 0.1423 = 0.8577
Method 2: z of 1.07 = 0.8577
or 85.77%

$z = \frac{x - \mu}{\sigma}$ $x = z\sigma + \mu$

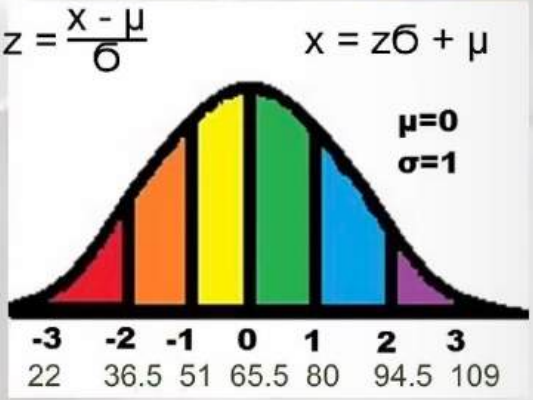


Harder Questions

$\mu = 65.5$
$\sigma = 14.5$

- What if you are looking at a probability more than 50%?
  - Such as the probability students will score *less than 90?*

$$z = \frac{x - \mu}{\sigma} \qquad x = z\sigma + \mu$$

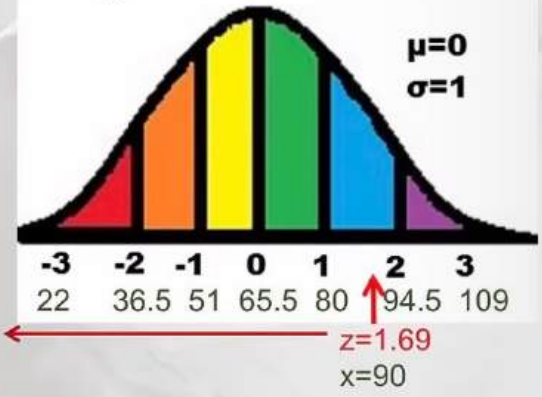


Harder Questions

$\mu = 65.5$
$\sigma = 14.5$

- What is the probability students will score *less than 90?*
- *For areas to the left of a specified z value, use the table entry directly.*

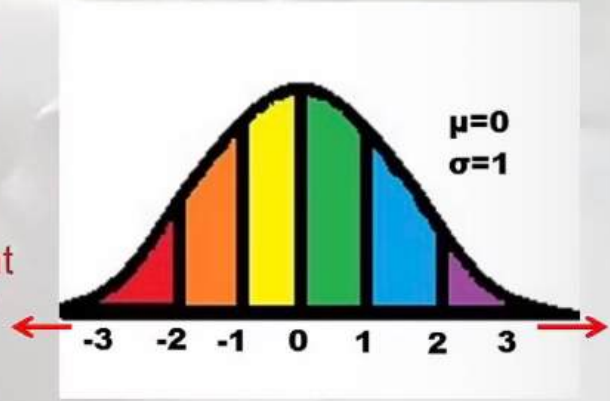
$$z = \frac{x - \mu}{\sigma} \qquad x = z\sigma + \mu$$

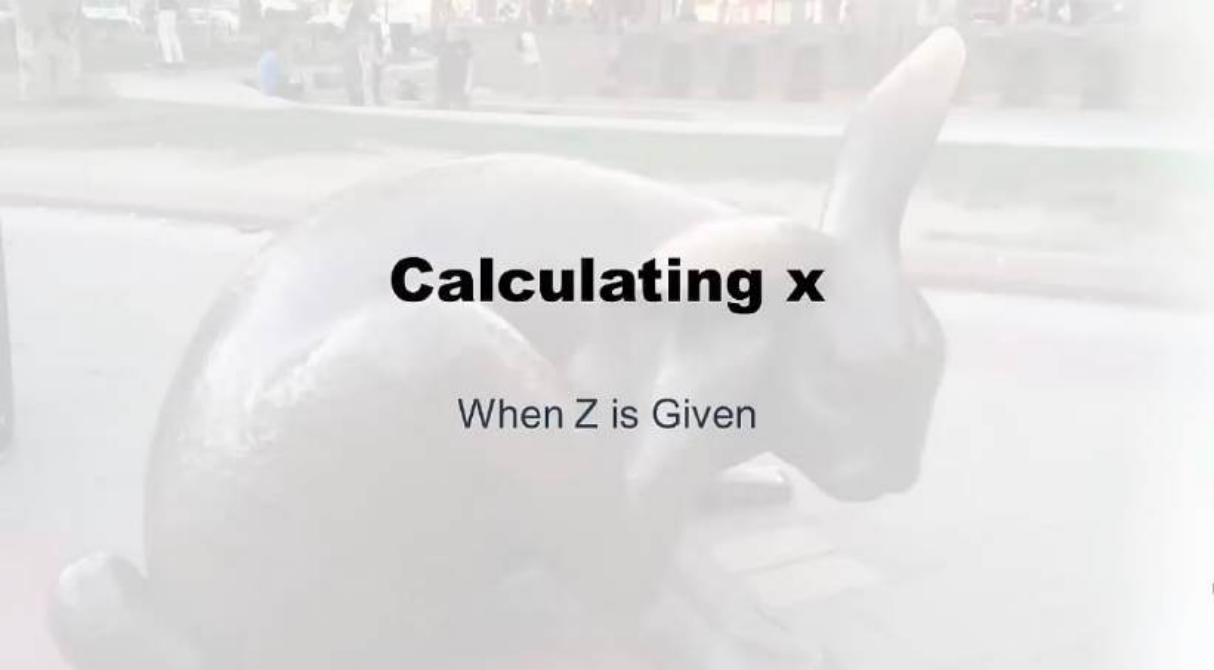


Probability for z=1.69 = 0.9545 , or 95.45%

Note About Z Table

1. Treat all areas (probabilities) to the left of z = -3.49 as p = 0.0000
2. Treat all areas (probabilities) to the right of z = 3.49 as p = 1.000





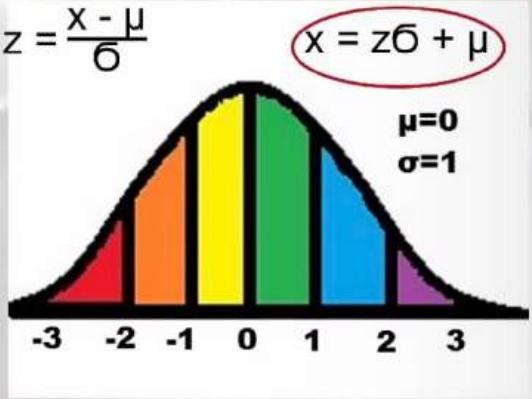
Calculating x

When Z is Given

Calculating x Questions

$\mu = 65.5$
$\sigma = 14.5$

- When calculating x, you need to be given μ and σ (just like when calculating z)
- But you also need to be either
  - Given a z score or
  - Given a probability so you can look up the z in the table
- Examples of each of these questions



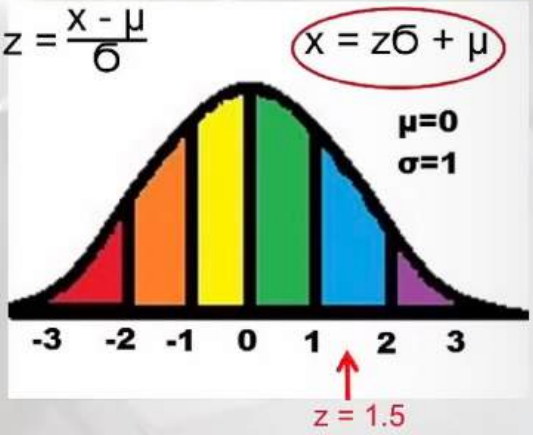
Calculating x Questions

$\mu = 65.5$
$\sigma = 14.5$

- What is the score on this distribution that is at z=1.5?

$$z = \frac{x - \mu}{\sigma}$$

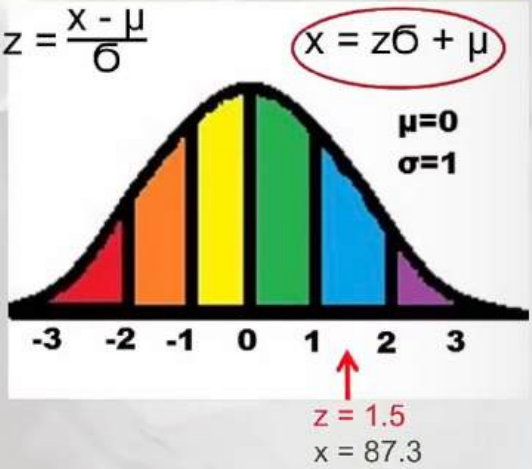
$$x = z\sigma + \mu$$



Calculating x Questions

$\mu = 65.5$
$\sigma = 14.5$

- What is the score on this distribution that is at z=1.5?
- x = (1.5*14.5) + 65.5 = 87.3
- The score (x) is 87.3.



Calculating x Questions

$\mu = 65.5$
$\sigma = 14.5$

- What is the score that marks the top 7% of the scores?
- We are looking for the z at p=0.0700

$$z = \frac{x - \mu}{\sigma}$$

$$x = z\sigma + \mu$$

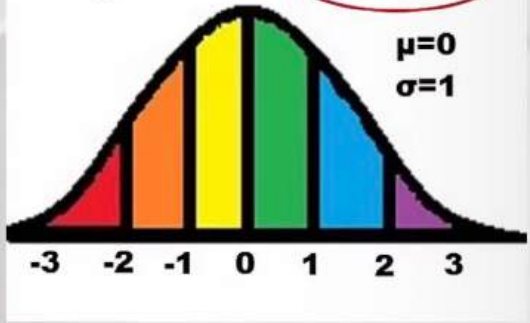


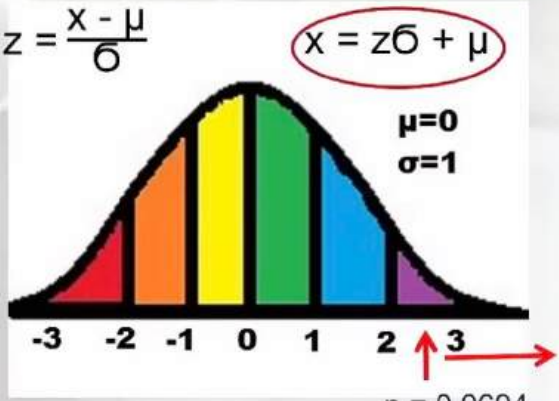
Table entry for z is the area to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451

Calculating x Questions

$\mu = 65.5$
$\sigma = 14.5$

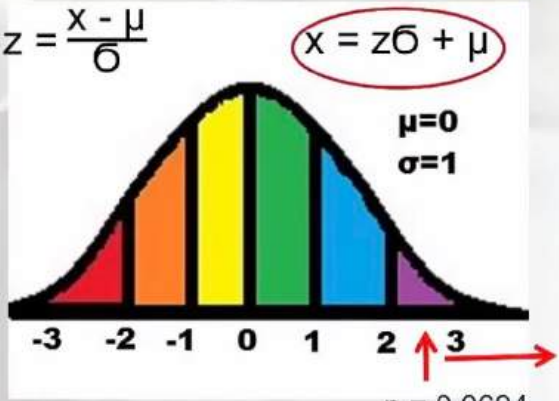
- What is the score that marks the top 7% of the scores?
- We are looking for the z at p=0.0700
  - Closest p in table is 0.0694
  - That maps to z = -1.48



Calculating x Questions

$\mu = 65.5$
$\sigma = 14.5$

- What is the score that marks the top 7% of the scores?
- We are looking for the z at p=0.0700
  - Closest p in table is 0.0694
  - That maps to z = -1.48
- Since we want "top 7%" we want positive z: z = 1.48



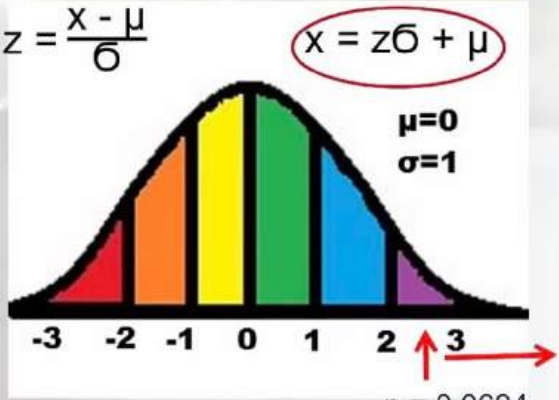
Calculating x Questions

$\mu = 65.5$
$\sigma = 14.5$

- What is the score that marks the top 7% of the scores?
- We are looking for the z at p=0.0700
  - Closest p in table is 0.0694
  - That maps to z = -1.48
- Since we want "top 7%" we want positive z: z = 1.48
- x = (1.48 * 14.5) + 65.5 = 87
- 87 is the score that marks the top 7% of scores

$z = \frac{x - \mu}{\sigma}$

$x = z\sigma + \mu$



p = 0.0694
z = 1.48
x = 87

Calculating x Questions

$\mu = 65.5$
$\sigma = 14.5$

- What is the score that marks the bottom 3% of the scores?
- We are looking for the z at p=0.0300

$$z = \frac{x - \mu}{\sigma}$$

$$x = z\sigma + \mu$$

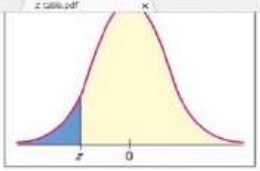


Table entry for z is the area to the left of z.

(a) Table of Areas to the Left of z

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823

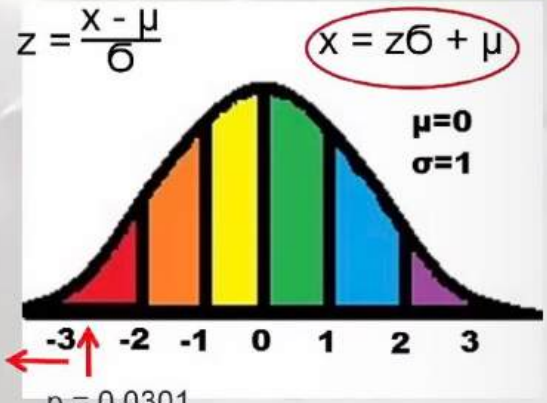
Calculating x Questions

$\mu = 65.5$
$\sigma = 14.5$

- What is the score that marks the bottom 3% of the scores?
- We are looking for the z at p=0.0300
  - Closest p in table is 0.0301
  - That maps to z = -1.88

$$z = \frac{x - \mu}{\sigma}$$

$$x = z\sigma + \mu$$



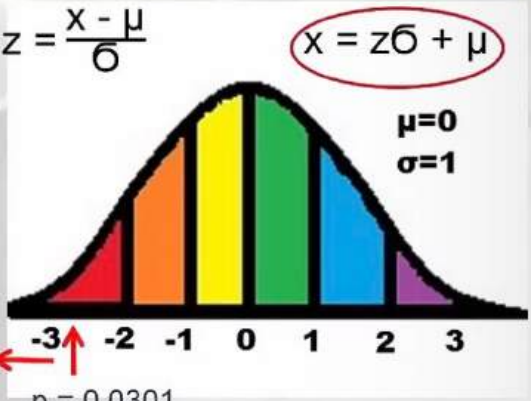
Calculating x Questions

$\mu = 65.5$
$\sigma = 14.5$

- What is the score that marks the bottom 3% of the scores?
- We are looking for the z at p=0.0300
  - Closest p in table is 0.0301
  - That maps to z = -1.88
- Since we want "bottom 3%" we keep the negative z = -1.88
- x = (-1.88 * 14.5) + 65.5 = 38.2
- 38.2 is the score that marks the bottom 3% of scores

$z = \frac{x - \mu}{\sigma}$

$x = z\sigma + \mu$



p = 0.0301
z = -1.88
x = 38.2

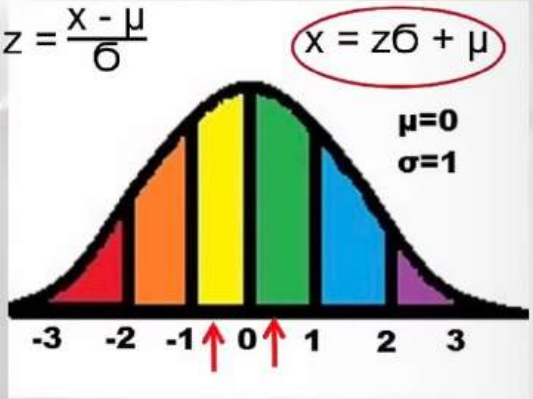
Calculating x Questions

$\mu = 65.5$
$\sigma = 14.5$

- What scores mark the middle 20% of the data?

$$z = \frac{x - \mu}{\sigma}$$

$$x = z\sigma + \mu$$



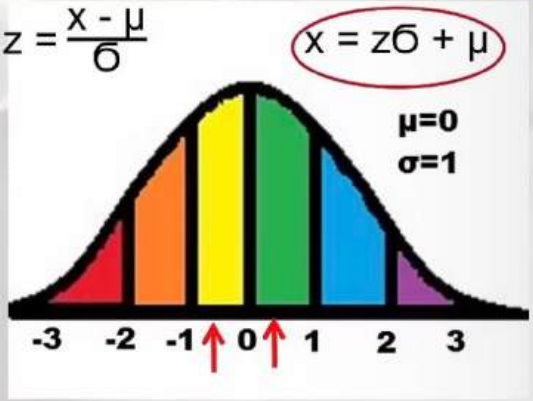
Calculating x Questions

$\mu = 65.5$
$\sigma = 14.5$

- What scores mark the middle 20% of the data?
- Strategy is to find the z-score for (1-0.2000)/2 = 0.4000

$$z = \frac{x - \mu}{\sigma}$$

$$x = z\sigma + \mu$$



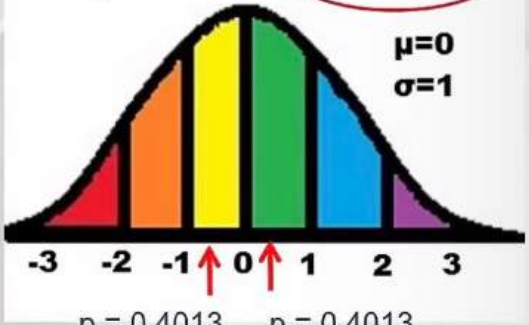
Calculating x Questions

$\mu = 65.5$
$\sigma = 14.5$

- What scores mark the middle 20% of the data?
- Strategy is to find the z-score for (1-0.2000)/2 = 0.4000
- For p = 0.4013, z = -0.25
- Also, z = 0.25 is on the positive side.

$$z = \frac{x - \mu}{\sigma}$$

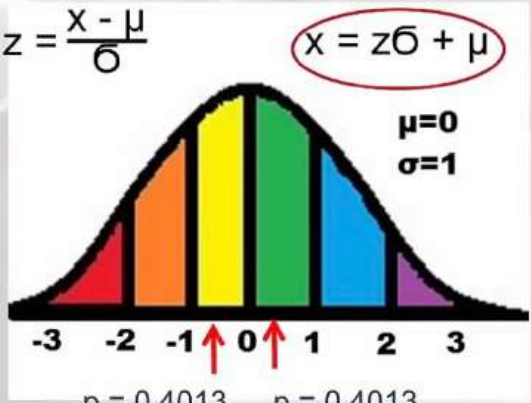
$$x = z\sigma + \mu$$

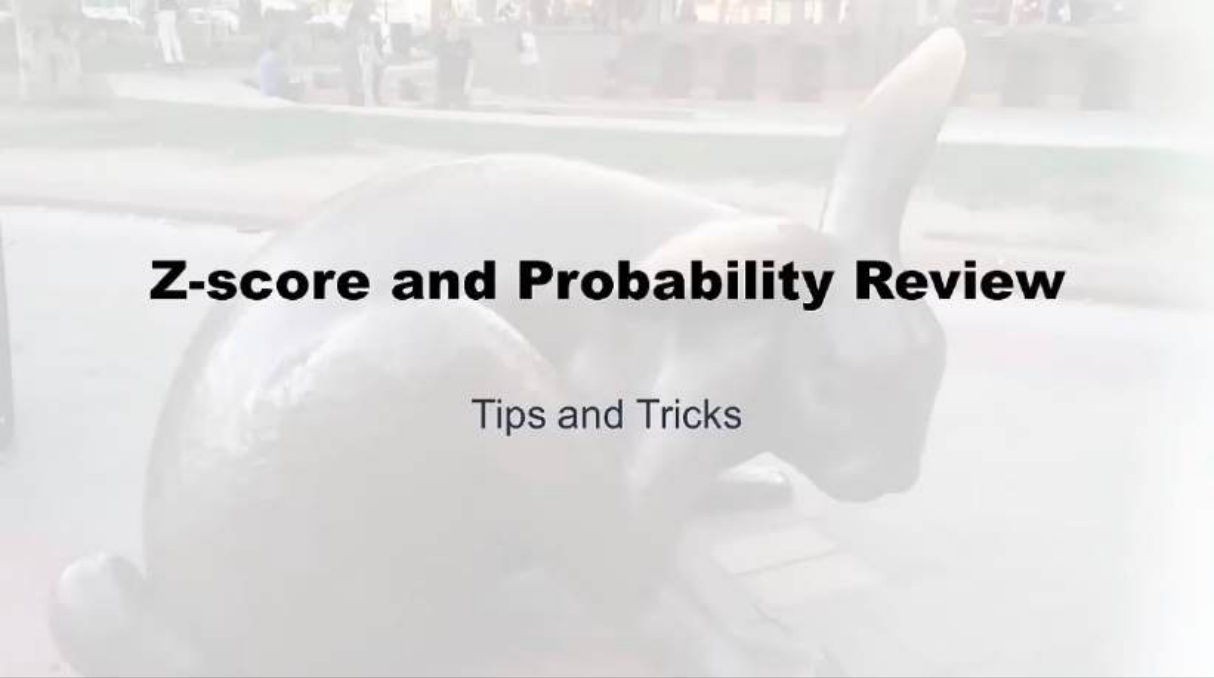


Calculating x Questions

$\mu = 65.5$
$\sigma = 14.5$

- What scores mark the middle 20% of the data?
- Strategy is to find the z-score for (1-0.2000)/2 = 0.4000
- For p = 0.4013, z = -0.25
- Also, z = 0.25 is on the positive side.
- x for the left side:
  - x = (-0.25*14.5) + 65.5 = 61.9
- x for the right side:
  - x = (0.25*14.5) + 65.5 = 69.1
- 61.9 and 69.1 mark the middle 20% of the data.





Z-score and Probability Review

Tips and Tricks

Z-Score Quiz

Question **Answer**

1. Where is x?

Z-Score Quiz

Question	Answer
1. Where is x?	1. Usually in the question.
2. What do you do with an x?	2. Calculate a z-score.
3. What do you do with a z?	3. Look it up in the Z table.
4. What if the question asks for x?	4. Use the x formula
5. What if the question gives you a p?	5. Dig around in the table to find the p to map back to z, then use x formula

Tips for Getting Z-Scores and Probabilities Right

1. Draw a picture: Graph out the question. Draw the curve, the line for μ, and where the x goes (above or below the μ).
   - If there is one x, shade in the part of the curve wanted (above or below).
   - If there are 2 x's, shade in the area wanted (usually in between them).
   - If it's a "calculate the x" question, put where the z or p is, and shade in the probability you are calculating.
2. x is usually in the question: The question must give you μ and σ, and students usually can find those, but then they can't find the x.
3. Don't mistake little z's for p's: Sometimes a little z-score (like 0.023) looks like a p. Don't be fooled! You still have to look it up.
4. Check logic against your picture: If you shaded in a big part of your picture, your probability should be bigger than 0.5000 or 50%.

Conclusion

- Introduction to the standard normal curve and z-score formulas
- How to calculate z-scores and look up probabilities
- How to calculate x if given a z-score or a probability

Ink on paper, courtesy of Online Collection of Brooklyn Museum

Chapter 7.4 & 7.5

Sampling Distributions and the Central Limit Theorem

Learning Objectives

At the end of this lecture, the student should be able to:

- State the statistical notation for parameters and statistics for two measures of variation.
- Name one type of inference and describe it.
- Explain the difference between a frequency distribution and sampling distribution
- Describe the Central Limit Theorem in either words or formulas.
- Describe how to calculate standard error

Introduction

- Parameters, Statistics, and Inferences
- Introduction to Sampling Distribution
- Central Limit Theorem
- Finding Probabilities Regarding X-bar

Photograph by David Hawgood

Parameters, Statistics, and Inferences

Review and Overview

Reminder of Statistic and Parameter

- A *statistic* is a numerical measure describing a *sample*.
- A *parameter* is a numerical measure describing a *population*.

Parameter

Statistic

Photographs by Che and Sandstein

Notation of Statistics and Parameters

Measure	Statistic	Parameter
Mean	$\overline{x}$ (x-bar)	μ (mu)
Variance	s^2	$б^2$ (sigma squared)
Standard Deviation	s	б (sigma)
Proportion	$\hat{p}$ (p-hat)	p

Types of Inferences

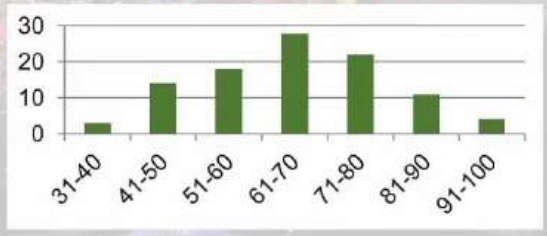
1. Estimation: we estimate the value of a population *parameter* using a *sample*
2. Testing: we do a test to help us make a *decision* about a population parameter
3. Regression: we make *predictions or forecasts* about a statistic

Requires understanding sampling distributions and the Central Limit Theorem

Frequency vs. Sampling Distribution

Frequency Distribution

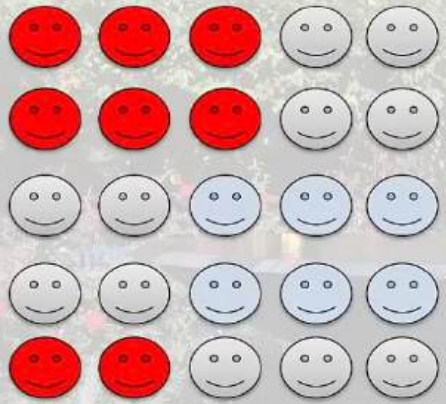
1. Make a histogram of a quantitative variable.
2. Draw the shape and name the distribution.



Sampling Distribution

1. Start with a population.
2. Decide on an n.
3. Take as many samples of n as possible from the population.
4. Make an x-bar for each sample.
5. Make a histogram of all the x-bars.

Explanation of Sampling Distribution

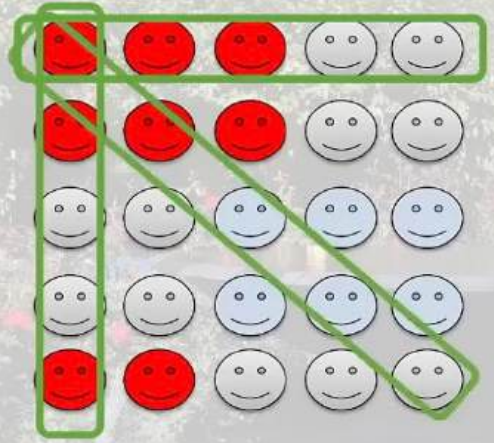


Sampling Distribution

1. Start with a population.
2. Decide on an n.
3. Take as many samples of n as possible from the population.
4. Make an x-bar for each sample.
5. Make a histogram of all the x-bars.

Explanation of Sampling Distribution

x-bar = 23



x-bar = 21

x-bar = 25

Sampling Distribution

1. Start with a population.
2. Decide on an n. **n=5**
3. Take as many samples of n as possible from the population.
4. Make an x-bar for each sample.
5. Make a histogram of all the x-bars.

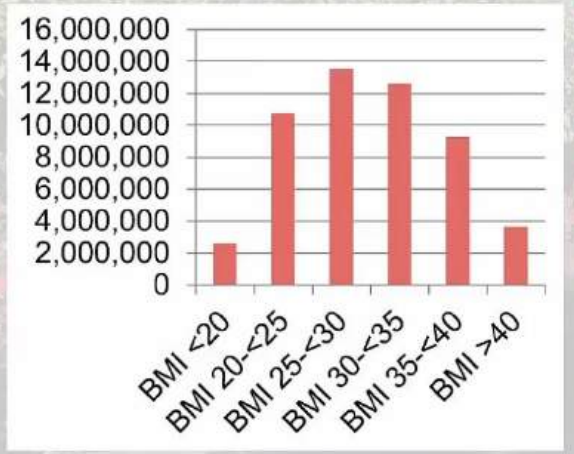
Explanation of Sampling Distribution

Class Limits of BMI	Frequency of x-bars
BMI <20	2,626,094
BMI 20-<25	10,758,762
BMI 25-<30	13,554,687
BMI 30-<35	12,605,250
BMI 35-<40	9,300,551
BMI >40	3,676,531
Total	52,521,875

Sampling Distribution

1. Start with a population.
2. Decide on an n. **n=5**
3. Take as many samples of n as possible from the population.
4. Make an x-bar for each sample.
5. Make a histogram of all the x-bars.

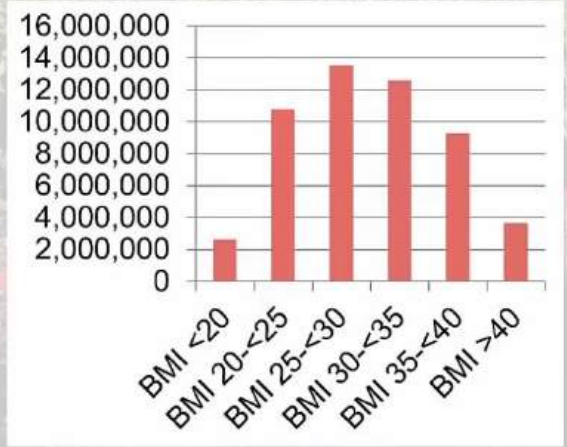
Explanation of Sampling Distribution



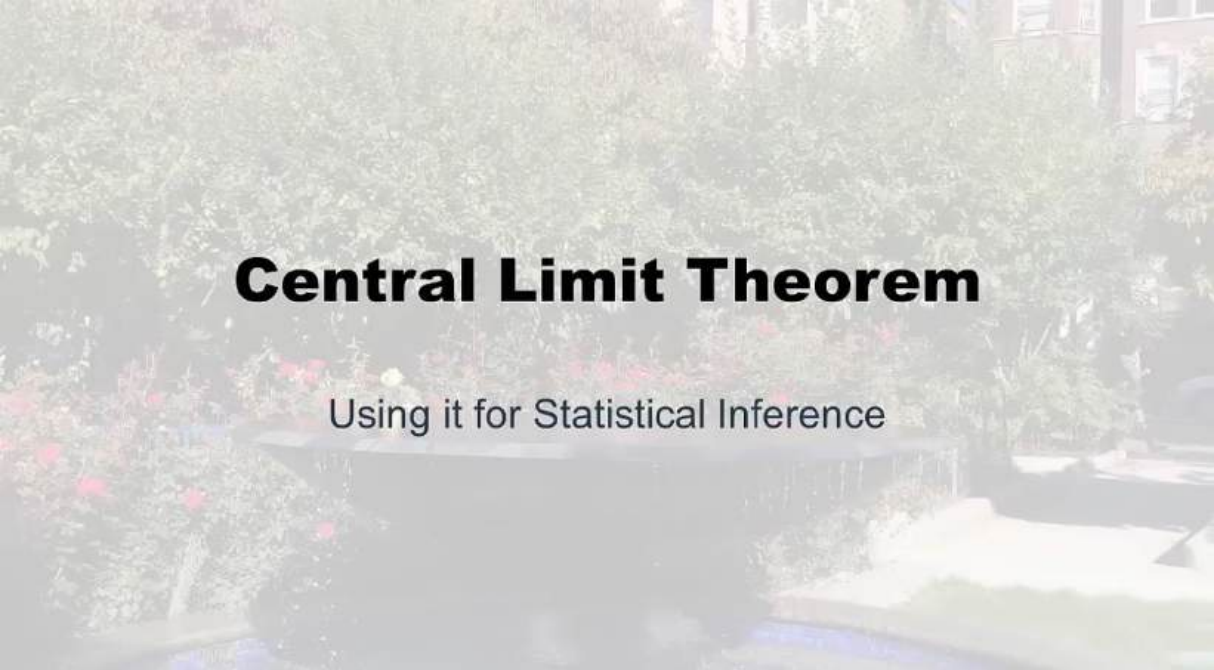
Sampling Distribution

1. Start with a population.
2. Decide on an n. **n=5**
3. Take as many samples of n as possible from the population.
4. Make an x-bar for each sample.
5. Make a histogram of all the x-bars.

Definition of a Sampling Distribution



- A sampling distribution is a probability distribution of a *sample statistic* based *on all possible simple random samples* of the *same size* from the *same population*.
- In the next section, we will talk about the Central Limit Theorem, which is a proof that shows how we can use a sampling distribution for inference.



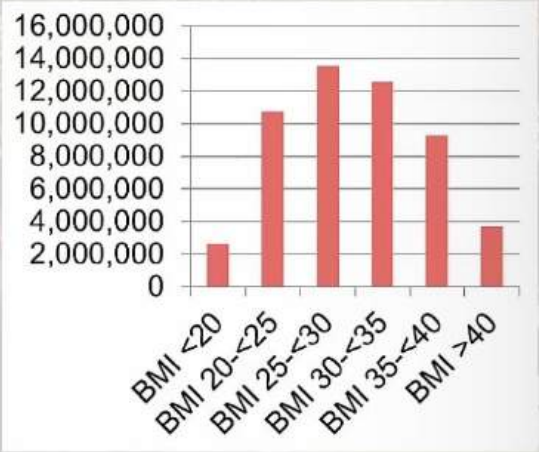
Central Limit Theorem

Using it for Statistical Inference

Central Limit Theorem: In Words

For any normal distribution:

1. The sampling distribution (the distributions of x-bars from all possible samples) is also a normal distribution
2. The mean of the x-bars is actually μ
3. The standard deviation of the x-bars is actually $\sigma/\sqrt{n}$



Central Limit Theorem: In Formulas

$$\mu_{x\text{-bars}} = \mu$$

$$\sigma_{x\text{-bars}} = \sigma/\sqrt{n}$$

$$z = \frac{x\text{-bar} - \mu_{x\text{-bars}}}{\sigma_{x\text{-bars}}} = \frac{x\text{-bar} - \mu}{\sigma/\sqrt{n}}$$

….where

- n is the sample size
- μ is the mean of the x distribution (population mean), and
- σ is the standard deviation of the x distribution (population standard deviation)

Central Limit Theorem: In Formulas

$\mu_{x\text{-bars}} = \mu$

$\sigma_{x\text{-bars}} = \sigma/\sqrt{n}$

$$z = \frac{\text{x-bar} - \mu_{x\text{-bars}}}{\sigma_{x\text{-bars}}} = \frac{\text{x-bar} - \mu}{\sigma/\sqrt{n}}$$

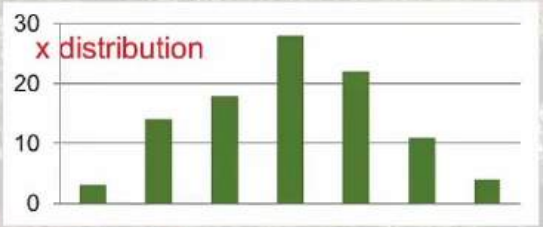
….where

- n is the sample size
  - Note: Only works when $n \geq 30$!
- μ is the mean of the x distribution (population mean), and
- σ is the standard deviation of the x distribution (population standard deviation)

The standard error is the standard deviation of the sampling distribution.
For the x-bar sampling distribution, standard error (SE) = $\sigma/\sqrt{n}$

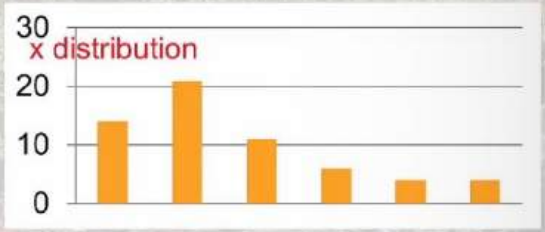
Central Limit Theorem

1. If the distribution of x is normal, then the distribution of x-bar is also normal.



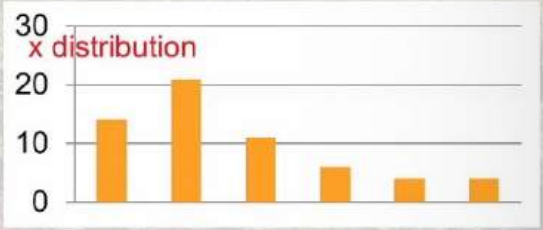
Central Limit Theorem

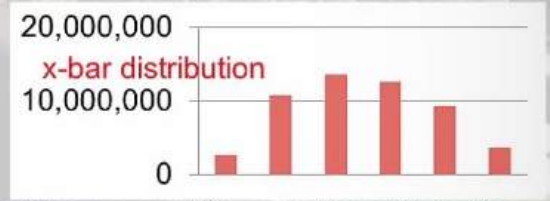
1. If the distribution of x is normal, then the distribution of x-bar is also normal.
2. Even if the distribution of x is NOT normal, as long as $n \geq 30$, the Central Limit Theorem says that the x-bar distribution is approximately normal.



Central Limit Theorem

1. If the distribution of x is normal, then the distribution of x-bar is also normal.
2. Even if the distribution of x is NOT normal, as long as $n \geq 30$, the Central Limit Theorem says that the x-bar distribution is approximately normal.





A sample statistic is considered ***unbiased*** if the mean of its sampling distribution equals the parameter being estimated.

Finding Probabilities Regarding X-bar

Applying the Central Limit Theorem

What Are We Doing?

Chapter 7.1-7.3

- We had a normally distributed x.
- We had a μ and a σ.
- We want to find the probability of selecting a value (from the population) above or below a value of x, so we use the z-score and z-table for probabilities.
- We used this formula: $z = \frac{x - \mu}{\sigma}$

Chapter 7.4-7.5

- We have a normally distributed x.
- We have a μ and a σ.
- We want to find the probability of selecting a *sample* n (from the population) with a mean value *(x-bar)* above or below a value of x-bar, so we use the z-score and z-table for probabilities.

We will use this formula: $z = \frac{\text{x-bar} - \mu}{\sigma/\sqrt{n}}$

Standard Error (SE)

What Are We Doing?

Chapter 7.1-7.3

- We had a normally distributed x.
- We had a μ and a σ.
- We want to find the probability of selecting a value (from the population) above or below a value of x, so we use the z-score and z-table for probabilities.
- We used this formula: $z = \frac{x - \mu}{\sigma}$

Chapter 7.4-7.5

- We have a normally distributed x.
- We have a μ and a σ.
- We want to find the probability of selecting a *sample* n (from the population) with a mean value *(x-bar)* above or below a value of x-bar, so we use the z-score and z-table for probabilities.

We will use this formula: $z = \frac{\text{x-bar} - \mu}{\sigma/\sqrt{n}}$

How to Find Probabilities Regarding X-Bar

1. Convert x-bar to a z-score using the following formula:

$$z = \frac{\text{x-bar} - \mu}{\sigma/\sqrt{n}}$$

2. Look up the probability for the z-score in the z-table (like in Chapters 7.2-7.3, only this is about x-bar).

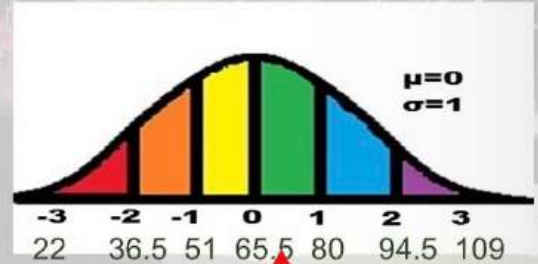
Remember the Students?

- Assume the 100-student class is a population.
- Now I have to pick an n
  - Let's pick 49.
- To pass the class, students have to get at least 70, which is a C.
- *Question:* What is the probability of me selecting a sample of 49 students with an x-bar greater than 70?

$$z = \frac{\text{x-bar} - \mu}{\sigma/\sqrt{n}}$$

μ = 65.5
σ = 14.5

x-bar = 70
n=49



Probability for x-bar will be smaller than for x

Remember the Students?

- Assume the 100-student class is a population.
- Now I have to pick an n
  - Let's pick 49.
- To pass the class, students have to get at least 70, which is a C.
- *Question:* What is the probability of me selecting a sample of 49 students with an x-bar greater than 70?

$$z = \frac{\text{x-bar} - \mu}{\sigma/\sqrt{n}}$$

$$\sigma/\sqrt{n} = SE$$

$$z = \frac{\text{x-bar} - \mu}{SE}$$

$\mu = 65.5$
$\sigma = 14.5$

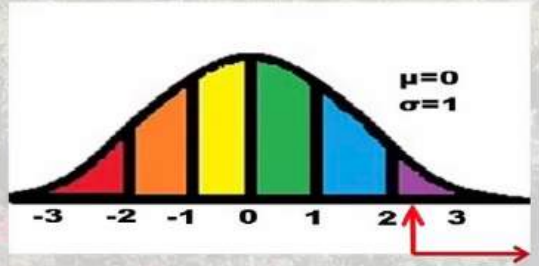
x-bar = 70
n=49

$$SE = \sigma/\sqrt{n} = 14.5/\sqrt{49} = 2.1$$

$$z = (70-65.5)/2.1 = 2.17$$

Remember the Students?

- Assume the 100-student class is a population.
- Now I have to pick an n
  - Let's pick 49.
- To pass the class, students have to get at least 70, which is a C.
- *Question:* What is the probability of me selecting a sample of 49 students with an x-bar greater than 70?



$z = (70-65.5)/2.1 = 2.17$

Use -2.17*

$p = 0.0150$

Remember the Students?

- Assume the 100-student class is a population.
- Now I have to pick an n
  - Let's pick 49.
- To pass the class, students have to get at least 70, which is a C.
- *Question:* What is the probability of me selecting a sample of 49 students with an x-bar greater than 70?

z = (70-65.5)/2.1 = 2.17
Use -2.17*
p = 0.0150

Probability is 0.0150, or 1.5%

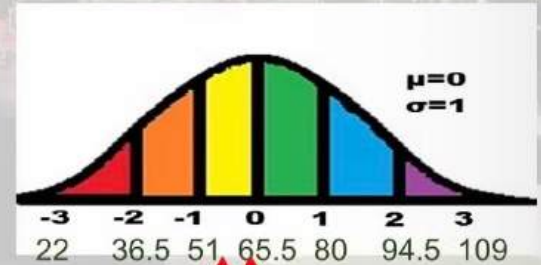
Remember the Students?

- Assume the 100-student class is a population.
- Now I have to pick an n
  - Let's pick 36.
- *Question:* What is the probability of me selecting a sample of 36 students with an x-bar between 60 and 65?

$$z = \frac{\text{x-bar} - \mu}{\sigma/\sqrt{n}}$$

μ = 65.5
σ = 14.5

x-bar1 = 60
x-bar2 = 65
n=36



Probability for x-bar will be smaller than for x

Remember the Students?

- Assume the 100-student class is a population.
- Now I have to pick an n
  - Let's pick 36.
- *Question:* What is the probability of me selecting a sample of 36 students with an x-bar between 60 and 65?

$$z = \frac{\text{x-bar} - \mu}{\sigma/\sqrt{n}}$$

$$\sigma/\sqrt{n} = SE$$

$$z = \frac{\text{x-bar} - \mu}{SE}$$

μ = 65.5
σ = 14.5

x-bar1 = 60
x-bar2 = 65
n=36

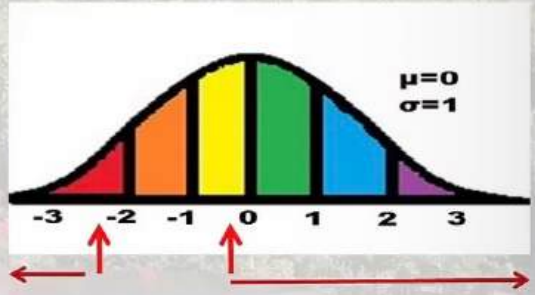
$SE = \sigma/\sqrt{n} = 14.5/\sqrt{36} = 2.4$

$z1 = (60-65.5)/2.4 = -2.28$

$z2 = (65-65.5)/2.4 = -0.21$

Remember the Students?

- Assume the 100-student class is a population.
- Now I have to pick an n
  - Let's pick 36.
- *Question:* What is the probability of me selecting a sample of 36 students with an x-bar between 60 and 65?



z1 = (60-65.5)/2.4 = -2.28

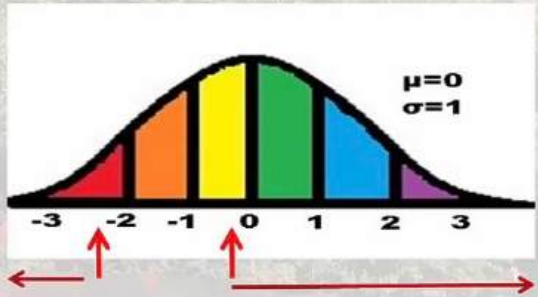
p1 = 0.0113

z2 = (65-65.5)/2.4 = -0.21

***Use 0.21 *** p2 = 0.5832

Remember the Students?

- Assume the 100-student class is a population.
- Now I have to pick an n
  - Let's pick 36.
- *Question:* What is the probability of me selecting a sample of 36 students with an x-bar between 60 and 65?



z1 = (60-65.5)/2.4 = -2.28

p1 = 0.0113

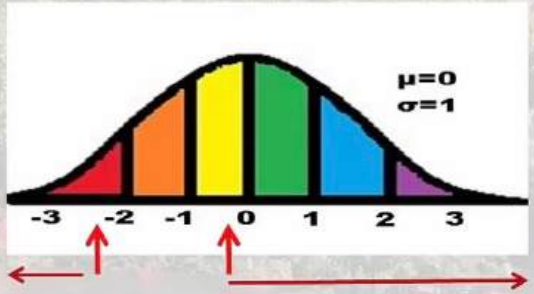
z2 = (65-65.5)/2.4 = -0.21

p2 = 0.5832

1 – 0.0113 – 0.5832 = 0.4055

Remember the Students?

- Assume the 100-student class is a population.
- Now I have to pick an n
  - Let's pick 36.
- *Question:* What is the probability of me selecting a sample of 36 students with an x-bar between 60 and 65?



z1 = (60-65.5)/2.4 = -2.28

p1 = 0.0113

z2 = (65-65.5)/2.4 = -0.21

p2 = 0.5832

The probability is 0.4055 or 41%

Conclusion

- Reviewed parameters and statistics, and discussed inferences
- Description of sampling distribution
- Presented Central Limit Theorem
- Examples of finding probabilities regarding x-bar

Photograph by Steve Cadman